Clausen

TR-01-265

WADC TECHNICAL REPORT 56-20

AN ELEMENTARY APPROACH TO THE ANALYSIS OF VARIANCE

M. ALEXANDER
~~PAUL R. RIDER~~
H. LEON HARTER
MARY D. LUM

AERONAUTICAL RESEARCH LABORATORY

FEBRUARY 1956

WRIGHT AIR DEVELOPMENT CENTER

NOTICES

- - - - -

Qualified requesters may obtain copies of this report from the ASTIA Document Service Center, Knott Building, Dayton 2, Ohio.

- - - - -

This report has been released to the Office of Technical Services, U. S. Department of Commerce, Washington 25, D. C., for sale to the general public.

- - - - -

Copies of WADC Technical Reports and Technical Notes should not be returned to the Wright Air Development Center unless return is required by security considerations, contractual obligations, or notice on a specific document.

WADC TECHNICAL REPORT 56-20

AD-093394

AN ELEMENTARY APPROACH TO THE ANALYSIS OF VARIANCE

PAUL R. RIDER

H. LEON HARTER

MARY D. LUM

FEBRUARY 1956

AERONAUTICAL RESEARCH LABORATORY
PROJECT 7060
TASK 70418

WRIGHT AIR DEVELOPMENT CENTER
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

Carpenter Litho & Prtg. Co., Springfield,O.
800 - May 1956

FOREWORD

This report was prepared for the Applied Mathematics Branch, Aeronautical Research Laboratory, Directorate of Research, Wright Air Development Center by Dr. Paul R. Rider, Dr. H. Leon Harter, and Mrs. Mary D. Lum under Task 70418, "Investigation of Analysis of Variance". The authors wish to thank Mrs. Geraldine K. Campbell for her excellent typing of the report, especially the difficult parts involving mathematical symbols and tabular matter.

The work was performed at the request of Colonel Aldro Lingard, Chief of the Aeronautical Research Laboratory, who expressed the opinion that there exists a need for an elementary exposition of the analysis of variance. The objective is to present this technique, a powerful statistical tool, in a manner that will be directly useful to engineers and other scientists engaged in research and development at Wright Air Development Center, as well as at other installations in the Department of Defense.

This is the fourth of a series of reports, a list of which follows.

SERIES OF REPORTS ON ANALYSIS OF VARIANCE

(1) Mentzer, E.G., Tests by the Analysis of Variance, WADC Technical Report 53-23.

(2) Harter, H. Leon; and Lum, Mary D., Partially Hierarchal Models in the Analysis of Variance, WADC Technical Report 55-33.

(3) (a) Bozivich, Helen; Bancroft, T.A.; Hartley, H.O.; and Huntsberger, David V., Analysis of Variance: Preliminary Tests, Pooling, and Linear Models, WADC Technical Report 55-244, Volume I, Preliminary Tests of Significance and Pooling Procedures for Certain Incompletely Specified Models.

(b) Wilk, M.B.; and Kempthorne, O., Analysis of Variance Preliminary Tests, Pooling, and Linear Models, WADC Technical Report 55-244, Volume II, Derived Linear Models and Their Use in the Analysis of Randomized Experiments.

(4) Rider, Paul R.; Harter, H. Leon; and Lum, Mary D., An Elementary Approach to the Analysis of Variance, WADC Technical Report 56-20.

ABSTRACT

An introduction to the Analysis of Variance is given. Several important experimental designs to which this statistical technique is applicable are discussed, as are multiple comparison tests which can be used after the analysis of variance has been made. Transformations employed prior to analysis are also treated. An extensive bibliography is to be found at the end of the report.

PUBLICATION REVIEW

This report has been reviewed and is approved.

FOR THE COMMANDER:

Aldro Lingard

ALDRO LINGARD
Colonel, USAF
Chief, Aeronautical Research Laboratory
Directorate of Research

TABLE OF CONTENTS

1. INTRODUCTION

The Analysis of Variance is a statistical technique which separates the variation in a set of data into parts which are linear combinations of estimates of the components of variance (see next section) due to different factors. After the variation has been so resolved, the process compares the part of the variation which includes the effect due to a certain factor (or combination of factors) with the part which excludes the effect due to that factor. If the former is of sufficiently greater magnitude than the latter, then one can conclude that the factor under consideration has contributed significantly to the variation. Thus, is this manner, the Analysis of Variance facilitates determining whether the factors under consideration have significantly influenced the variation in the data.

2. DEFINITIONS

In this section are given some definitions and certain symbols and notations which will be used throughout this report.

Let $x_1, x_2, \cdots, x_N$ be a set of N values of a variable x. Then the mean (more explicitly the arithmetic mean) of this set is given by

$$\overline{x} = (1/N)\Sigma x, \tag{1}$$

where Σx indicates $x_1 + x_2 + \cdots + x_N$, that is, the sum of the x's.

A measure of variability is provided by the variance of the x's, which is defined as

$$s^2 = (1/n)\Sigma(x-\overline{x})^2 \ , \ n = N-1. \tag{2}$$

The quantity n is the number of degrees of freedom.

Here it is one less than the number of x's.

Example. $x_1 = 4, \ x_2 = 7, \ x_3 = 8, \ x_4 = 5.$

$$\overline{x} = (1/4)(4 + 7 + 8 + 5) = 6.$$

$$s^2 = \frac{1}{4-1}[(4-6)^2 + (7-6)^2 + (8-6)^2 + (5-6)^2] = 10/3.$$

The variance is ordinarily more easily calculated by a formula equivalent to (2), namely,

$$s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/N}{n}. \qquad (3)$$

For the above example,

$$\Sigma x^2 = 4^2 + 7^2 + 8^2 + 5^2 = 154,$$

and (3) gives

$$s^2 = \frac{154-(24)^2/4}{4-1} = 10/3,$$

as before.

Although in this example formula (3) does not shorten the calculation of s^2, nevertheless, in general it will be found more convenient than (2).

It is appropriate at this point to define the standard deviation; it is s, the square root of the variance.

3. DECOMPOSITION OF THE VARIANCE

As a simple example of the analysis of variance consider the data in Table 3.1. It might be imagined that these data, which actually are fictitious, represent the number of hours required before failure, for certain items that are being tested. Suppose that the figures in a

Table 3.1

	(1)	(2)	(3)	
	5	6	10	
	6	6	9	
	4	3	8	
	5	5	5	
Total	20	20	32	72
Mean	5	5	8	6

given column are for items supplied by a specified factory. The question which is to be answered is the following: Are the products supplied by the three factories significantly different, that is, is the variation among the means of the three columns greater than that which would be expected to occur as a matter of chance?

The first important point to be noted is that the total variation can be resolved into the variation within the columns and the variation of the column means about the general mean. It is a matter of elementary algebra to prove that the sum of squares of deviations of all of the values from the general mean is equal to the sum of squares of deviations of their values from their respective column means plus the (weighted) sum of squares of deviations of the column means from the general mean. (The weight used for a column mean is the number of values in the column. These numbers are not necessarily equal. Here the weight of each column mean is four.) We shall not prove this statement here but will verify it numerically for the present example.

The sum of squares of deviations from the general mean is

$$(5-6)^2 + (6-6)^2 + (10-6)^2 + (6-6)^2 + (6-6)^2 + (9-6)^2$$
$$+(4-6)^2 + (3-6)^2 + (8-6)^2 + (5-6)^2 + (5-6)^2 + (5-6)^2$$
$$= 1 + 0 + 16 + 0 + 0 + 9 + 4 + 9 + 4 + 1 + 1 + 1 = 46.$$

The sum of squares of deviations from the column means is

$$(5-5)^2 + (6-5)^2 + (4-5)^2 + (5-5)^2 \qquad \text{Column (1)}$$
$$+(6-5)^2 + (6-5)^2 +(3-5)^2 + (5-5)^2 \qquad \text{Column (2)}$$
$$+(10-8)^2 + (9-8)^2 + (8-8)^2 + (5-8)^2 \qquad \text{Column (3)}$$
$$= 0 + 1 + 1 + 0 \Big| + 1 + 1 + 4 + 0 \Big| + 4 + 1 + 0 + 9 \Big| = 22.$$

The weighted sum of squares of the column means from the general mean is

$$4[(5-6)^2 + (5-6)^2 + (8-6)^2] = 4(1 + 1 + 4) = 24.$$

As a check we note that 22 + 24 = 46.

In practice these sums of squares of deviations, usually called "sums of squares" for short, will be calculated more easily by using the formula

$$\Sigma x^2 - (\Sigma x)^2/N, \tag{4}$$

which is the numerator of (3). Thus, the total sum of squares is

$$5^2 + 6^2 + 10^2 + 6^2 + 6^2 + 9^2 + 4^2 + 3^2 + 8^2 + 5^2 + 5^2 + 5^2 - (72)^2/12$$

$$= 478 - 432 = 46.$$

The sum of squares of column means is

$$4(5^2 + 5^2 + 8^2) - (72)^2/12 = 24.$$

Alternatively this can be calculated by using the totals of columns and the factor 1/4 instead of 4, thus:

$$\frac{1}{4}[(20)^2 + (20)^2 + (32)^2] - (72)^2/12 = 24.$$

The within-columns sum of squares is calculated by subtraction: 46-24 = 22.

4. TESTING FOR SIGNIFICANCE

It is assumed that each group of data in Table 3. 1 constitutes a random sample from an infinitely large set of values of hours-to-failure of items of the type being tested. This infinitely large set of values is referred to as the population. The variance of the population is denoted by σ^2.

An estimate of σ^2 can be obtained in several ways. The total sum of squares, 46, can be divided by the number of degrees of freedom, 12-1=11. Or, the column sum of squares, 24, can be divided by the appropriate number of degrees of freedom, which is one less than the number of columns, that is, 3-1=2. A third estimate can be obtained by dividing the within-columns sum of squares by the appropriate number of degrees of freedom. There are 4 values in each column, consequently 4-1=3 degrees of freedom per column, or a total of

3 x 3 = 9 degrees of freedom. Each of these estimates is called a mean square. It is the second and third estimates which are of greatest interest. Results are summarized in Table 4. 1.

Table 4. 1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Column means	24	3-1 = 2	24/2 = 12
Within columns	22	3(4-1) = 9	22/9 = 2.44
Total	46	12-1 = 11	

It will be noted that the sum of squares (as previously stated) and also the degrees of freedom, but not the mean squares, are additive. Thus, 24 + 22 = 46 and 2 + 9 = 11.

Now let it be assumed that the variability of the product is the same in all of the factories, in other words, that the three columns in Table 3. 1 are random samples from populations with equal variances. Then a test can be made of the hypothesis that the variance of the column means is zero, in other words that the population column means are equal, which would mean in the present example that the average time to failure is the same for all three factories. (This is sometimes called the null hypothesis.) If this hypothesis is true, then the mean square for column means should not be too different from the within-columns mean square. If it is too much greater the hypothesis is contradicted and it may be concluded that there is more of a difference in the products put out by the three factories than can reasonably be attributed to chance. One then states that the difference is significant (in the statistical sense). The reader should be cautioned that a significant difference does not necessarily imply that the actual magnitude of the difference is large. On the contrary, the magnitude may be extremely small. It does indicate that the difference is of a causal nature and cannot be attributed to chance.

To test the hypothesis one forms the ratio (sometimes called the variance ratio)

$$F = 12/2.44 = 4.92.$$

The mean square in the numerator of the ratio is based upon 2 degrees of freedom, that in the denominator upon 9 degrees of freedom. There are tables* which give the values of F that will be exceeded a certain per cent of the time. Reference to such tables shows that for 2 degrees of freedom in the numerator and 9 in the denominator, the value of F which will be exceeded 5 per cent of the time is $F_{.05} = 4.26$. A value of F which exceeds $F_{.05}$ is said to be significant at the 5 per cent level. Similarly, the value of F which will be exceeded 1 per cent of the time is $F_{.01} = 8.02$. A value of F which exceeds this is said to be significant at the 1 per cent level. The value of F for this example is thus significant at the 5 per cent level but not at the 1 per cent level. The conclusion is that the null hypothesis is contradicted and that there is some reason to suppose that there is a difference among the overall average qualities of the products of the three factories, and not merely among the average qualities of the samples included in the experiment.

5. TWO-WAY CLASSIFICATION

Let it now be supposed that the data of Table 1 are classified according to rows as well as columns. Thus, each row might indicate an individual testing machine, so that, for example, the number 3 in column 2 and row 3 is the number of hours that were required for the item from factory 2 to fail on testing machine number 3.

The totals and means of the rows, as well as the totals and the means of columns, are shown in Table 5. 1.

Table 5. 1

	(1)	(2)	(3)	Total	Mean
(1)	5	6	10	21	7
(2)	6	6	9	21	7
(3)	4	3	8	15	5
(4)	5	5	5	15	5
Total	20	20	32	72	24
Mean	5	5	8	18	6

*See, for example, Snedecor, George W., Statistical Methods.

The weighted sum of squares of deviations of the row means from the general mean is

$$3[(7-6)^2 + (7-6)^2 + (5-6)^2 + (5-6)^2] = 12.$$

The weight 3 preceding the brackets is the number of values in each row. This sum of squares can also be calculated as

$$3(7^2 + 7^2 + 5^2 + 5^2) - (72)^2/12 = 12$$

or, by using row totals instead of row means, as

$$\frac{1}{3}(21^2 + 21^2 + 15^2 + 15^2) - (72)^2/12 = 12.$$

(Cf. preceding section.)

The sum of squares of deviations for column means plus that for row means is 12 + 24 = 36. When this value is subtracted from the total sum of squares, namely 46, there is a residual of 10, which will be referred to as the discrepance. It is composed of interaction and error. The interaction is that part of the variation which is caused by the interplay of the factors at work; for example there might be a tendency of the product of a certain factory to test higher on a certain machine. The error is that part of the variation which may be regarded as due to chance. There is a formula by which the discrepance may be calculated directly, but it is easiest and simplest to obtain it by subtraction.

Table 5.2 can now be constructed. It is like Table 4.1 except that the analysis of variance has been carried further. Note that the number of degrees of freedom for discrepance can be obtained by multiplying the degrees of freedom for columns and for rows.

Table 5.2

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Column means	24	3-1 = 2	24/2 = 12
Row means	12	4-1 = 3	12/3 = 4
Discrepance	10	(3-1)(4-1) = 6	10/6 = 1.67
Total	46	12-1 = 11	

Significance tests can now be made only if it can be assumed that there is no interaction. Possibly there is available some previous information on this point and it is believed that this is a reasonable assumption. If so, the discrepance mean square is used as a standard of comparison.

For columns,

$$F = 12/1.67 = 7.19.$$

Tables show that for 2 degrees of freedom in the numerator and 6 in the denominator, $F_{.05} = 5.14$ and $F_{.01} = 10.92$. The value of 7.19 is thus significant at a level somewhere between the 5 per cent and the 1 per cent levels. It is higher above the 5 per cent level than before, strengthening the conclusion that the products of the three factories are different. In general, one should note that if the row effects are analyzed out, the mean square of the comparison term might be so decreased (and the F-ratio consequently so increased) as to detect a significant difference in the columns which had not been apparent before.

For rows,

$$F = 4/1.67 = 2.40.$$

Since, for the degrees of freedom 3 and 6, tables show $F_{.05} = 4.76$ and $F_{.01} = 9.78$, it may not be concluded that there is a row effect, that is, a difference in the testing machines .

6. REPLICATIONS

In the statistical experiment described in the preceding sections it would have been desirable to have more than one item tested from each factory by each machine. The number of items so tested is referred to as the number of replications; thus, if five items from each factory are tested on each machine we say that there are five replications. Replications not only yield more accurate results (because the sample is larger), but they provide a means of testing the significance of the interaction, which otherwise is impossible. The variation in the individual values in the various classes can be used as a standard of comparison.

Of course in the example cited above there would still be variation among individual items, since a test-to-failure is destructive, so perhaps a more appropriate example would have been one in which several measurements of hardness are made on items from different factories or sources of supply by different persons or with different instruments.

As a numerical example consider Table 6. 1. The data therein can be analyzed as were the data of Table 5. 1.

Table 6. 1

				Total	Mean
	5	6	10		
	4	6	11		
	9	12	21	42	7
	6	6	9		
	5	4	6		
	11	10	15	36	6
	4	3	8		
	7	4	10		
	11	7	18	36	6
	5	5	5		
	4	6	5		
	9	11	10	30	5
Total	40	40	64	144	24
Mean	5	5	8	18	6

For the total sum of squares of the twelve averages $9/2 = 4.5$, $12/2 = 6$, $21/2 = 10.5$, etc., is found

$$2[(4.5-6)^2 + (6-6)^2 + (10.5-6)^2 + (5.5-6)^2 + (5.6)^2 + (7.5-6)^2 + (5.5-6)^2 + (3.5-6)^2 + (9-6)^2 + (4.5-6)^2 + (5.5-6)^2 + (5-6)^2]$$

$$= 2(2.25 + 0 + 20.25 + 0.25 + 1 + 2.25 + 0.25 + 6.25 + 9 + 2.25 + 0.25 + 1) = 2(45) = 90.$$

Short-cut computation:

$$\frac{1}{2}(9^2 + 12^2 + 21^2 + 11^2 + 10^2 + 15^2 + 11^2 + 7^2 + 18^2$$

$$+ 9^2 + 11^2 + 10^2) - (144)^2/24$$

$$= \frac{1}{2}(1908) - 864 = 90.$$

For the column sum of squares is obtained

$$8[(5-6)^2 + (5-6)^2 + (8-6)^2] = 48.$$

Short-cut computation:

$$\frac{1}{8}(40^2 + 40^2 + 64^2) - (144)^2/24$$

$$= \frac{1}{8}(1600 + 1600 + 4096) - 864 = 48.$$

Similarly, for rows,

$$6[(7-6)^2 + (6-6)^2 + (6-6)^2 + (5-6)^2] = 12.$$

Short-cut computation:

$$\frac{1}{6}(42^2 + 36^2 + 36^2 + 30^2) - (144)^2/24$$

$$= \frac{1}{6}(1764 + 1296 + 1296 + 900) - 864$$

$$= \frac{1}{6}(5256) - 864 = 876 - 864 = 12.$$

The interaction sum of squares is

$$90 - 48 - 12 = 30$$

So far use has not been made of the individual values. The total sum of squares if they are taken into consideration is

$$(5-6)^2 + (4-6)^2 + (6-6)^2 + (6-6)^2 + (10-6)^2 + (11-6)^2$$
$$+(6-6)^2 + (5-6)^2 + (6-6)^2 + (4-6)^2 + (9-6)^2 + (6-6)^2$$
$$+(4-6)^2 + (7-6)^2 + (3-6)^2 + (4-6)^2 + (8-6)^2 + (10-6)^2$$
$$+(5-6)^2 + (4-6)^2 + (5-6)^2 + (6-6)^2 + (5-6)^2 + (5-6)^2$$

$$= 1 + 4 + 0 + 0 + 16 + 25 + 0 + 1 + 0 + 4 + 9 + 0$$
$$+ 4 + 1 + 9 + 4 + 4 + 16 + 1 + 4 + 1 + 0 + 1 + 1 \quad = \quad 106.$$

Short-cut computation:

$$5^2 + 4^2 + 6^2 + 6^2 + 10^2 + 11^2 + 6^2 + 5^2 + 6^2 + 4^2 + 9^2 + 6^2$$
$$+4^2 + 7^2 + 3^2 + 4^2 + 8^2 + 10^2 + 5^2 + 4^2 + 5^2 + 6^2 + 5^2 + 5^2$$
$$-(144)^2/24 \qquad = \quad 106.$$

There is still a sum of squares unaccounted for:

$$106 - 48 - 12 - 30 = 16.$$

Actually this is the sum of squares of the deviations of the individual values from the means of their respective classes. This is sometimes termed the error sum of squares. It is

$$(5-4.5)^2 + (4-4.5)^2 + (6-6)^2 + (6-6)^2 + (10-10.5)^2 + (11-10.5)^2$$
$$+(6-5.5)^2 + (5-5.5)^2 + (6-5)^2 + (4-5)^2 + (9-7.5)^2 + (6-7.5)^2$$
$$+(4-5.5)^2 + (7-5.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (8-9)^2 + (10-9)^2$$
$$+(5-4.5)^2 + (4-4.5)^2 + (5-5.5)^2 + (6-5.5)^2 + (5-5)^2 + (5-5)^2$$

$$= 0.25 + 0.25 + 0 + 0 + 0.25 + 0.25 + 0.25 + 0.25 + 1 + 1$$
$$+ 2.25 + 2.25$$

$$+ 2.25 + 2.25 + 0.25 + 0.25 + 1 + 1 + 0.25 + 0.25 + 0.25$$
$$+ 0 + 0 \qquad = 16.$$

These results have been placed in an analysis of variance table, Table 6.2, in which the following abbreviations are employed.

SS	=	sum of squares,
DF	=	degrees of freedom,
MS	=	mean square.

Table 6.2

Source	SS	DF	MS
Columns	48	3-1 = 2	48/2 = 24
Rows	12	4-1 = 3	12/3 = 4
Interaction	30	(3-1)(4-1) = 6	30/6 = 5
Error	16	12(2-1) = 12	16/12 = 1.33
Total	106	24 - 1 = 23	

If now the error mean square is used as the denominator or comparison term in forming F-ratios, the results listed in Table 6.3 are obtained.

Table 6.3

Source	F (= Variance Ratio)	$F_{.05}$	$F_{.01}$
Interaction	5/1.33 = 3.75	3.00	4.82
Columns	24/1.33 = 18.0	3.88	6.93
Rows	4/1.33 = 3.0	3.49	5.95

The F-ratio for interaction is significant at the 5 per cent level (but not at the 1 per cent level), which would contradict the hypothesis that the effect of each machine is the same for all factories (and vice versa).

The F-ratio for columns is now over two and one-half times the 1 per cent value, and the hypothesis of no difference between factories is much more strongly refuted than before. The separating out of the interaction has thus provided a still more sensitive test.

The F-ratio for rows is not significant.

It has been tacitly assumed in the above example that both factories and machines are fixed factors (see next section).

7. FIXED AND RANDOM FACTORS

If the levels of a factor which are included in an experiment are considered to constitute the entire population of levels of that factor, the factor is said to be fixed. If, on the other hand, the levels included in the experiment represent a random sample from an infinite population of such levels, the factor is said to be random. If the levels included in the experiment represent a random sample from a larger (but finite) population, the factor is said to be semi-random. Only fixed and random factors will be discussed in this report.

An experimental model in which all of the factors are fixed is called a fixed model. A model in which all of the factors are random is called a random model. A mixed model is one in which both fixed and random factors are included.

For a fixed factorial model (all models considered so far have been factorial; see section 8 for definition), all main effects and interactions in a replicated experiment are tested by using the mean square for error as the denominator of the F-ratio. The reader will recall that this was done in the example of section 6.

If, on the other hand, one of the factors had been random, the other factor in this mixed model would have been tested by using the mean square for interaction as the denominator of the F-ratio. In the example of section 6, the experimenter was interested in the results of testing the product of three factories, using a particular set of four machines, and the effect of factories was tested by error. Suppose now the four machines represent a random sample from a much larger population of machines. Strictly speaking, if the population is finite, machines should be considered to be semi-random, but if the population size is quite large with respect to the sample size, it can be considered to be infinite for all practical purposes, and machines can be considered random. Now the effect of factories will be tested by interaction; that is, the F-ratio for factories will be 24/5 = 4. 8. For 2 and 6 degrees of freedom, $F_{.05} = 5.14$. Hence the effect of factories, which was found to be significant at the 1 per cent level when machines were considered to be fixed, is not significant even at the 5 per cent level when machines are considered to be random.

To give these results a practical interpretation, one can be quite sure that if all the product of the three factories were tested on these same four machines, a difference in overall factory means would be found. However, if the tests were made on an independent random sample of four machines, one cannot predict what the results would be.

If both factories and machines had been random in the example of section 6, both factors in this random model would have been tested by interaction. Of course testing machines by interaction would have made no difference, since the effect of machines is not significant when tested either by interaction or by error.

8. FACTORIAL, HIERARCHAL AND PARTIALLY HIERARCHAL MODELS

If the chosen levels of the various factors in an experiment are tried in all combinations, the experimental model is said to be factorial. As pointed out in section 7, all of the models considered so far have been of this type. If a different set of levels of a second factor is used for each level of a first factor, then the second factor is said to nest within the first factor, and the model is said to be hierarchal.

As an example of a two-factor hierarchal model, consider an experiment designed to study the maximum true air speed (in horizontal flight at low altitude) of planes of different types. Since each plane is of a definite type, planes are said to nest within aircraft types. Since only particular aircraft types are involved, aircraft types will be fixed. Since one is interested not in specific planes, but in all planes of the given types, the planes which will represent a given aircraft type should be chosen at random from the population of planes of that type. (Certain difficulties arise in the case of fixed nesting factors, which will not be discussed here.) Replication is needed to make possible a test of the significance of planes within aircraft types. The data in Table 8. 1 represent the maximum true air speed on three passes of two planes each of two aircraft types.

The total sum of squares is computed as follows (short-cut method):

$$(450)^2 + (460)^2 + (440)^2 + (400)^2 + (430)^2 + (400)^2 + (520)^2$$
$$+ (560)^2 + (540)^2 + (570)^2 + (590)^2 + (580)^2 - \frac{(5940)^2}{12}$$
$$= 57300 .$$

Table 8. 1

Aircraft Type 1		Aircraft Type 2	
Plane A	Plane B	Plane C	Plane D
450	400	520	570
460	430	560	590
440	400	540	580
1350	1230	1620	1740
2580		3360	

The short-cut computation for the sum of squares for aircraft types is:

$$[(2580)^2 + (3360)^2]/6 - (5940)^2/12 = 50700.$$

The sum of squares for planes within type 1 is

$$[(1350)^2 + (1230)^2]/3 - (2580)^2/6 = 2400.$$

The sum of squares for planes within type 2 is

$$[(1620)^2 + (1740)^2]/3 - (3360)^2/6 = 2400.$$

Hence the sum of squares for planes within types is

$$2400 + 2400 = 4800.$$

The sum of squares for error is then found by subtraction to be 57300 - 50700 - 4800 = 1800.

Since there are two aircraft types, the number of degrees of freedom for aircraft types is 2-1 = 1. Since there are two planes within each of two aircraft types, the number of degrees of freedom for planes within aircraft types is 2(2-1) = 2(1) = 2. Since there are three observations for each of four planes, there are 4(3-1) = 4(2) = 8 degrees of freedom for error. Since there are 12 observations in all, the total number of degrees of freedom is 12-1 = 11. The same result is obtained by adding the degrees of freedom for the various effects, thus: 1 + 2 + 8 = 11.

These results, together with the mean squares, are summarized in Table 8. 2.

In a hierarchal model such as this, the mean square for the nesting factor is used as the denominator of the F-ratio for testing the significance of the factor in which it nests. The nesting factor itself is

Table 8.2

Source of Variation	SS	DF	MS
Aircraft types	50,700	2-1 = 1	50,700/1 = 50,700
Planes within types	4,800	2(2-1) = 2	4,800/2 = 2,400
Error	1,800	4(3-1) = 8	1,800/8 = 225
Total	57,300	12-1 = 11	

tested by error, as shown in Table 8.3.

Table 8.3

Source of Variation	F (= Variance Ratio)	$F_{.05}$	$F_{.01}$
Aircraft types	50,700/2,400 = 21.12	18.51	98.49
Planes within types	2,400/ 225 = 10.67	4.46	8.65

The reader will note that, although the F-ratio for aircraft types is nearly twice that for planes within types, the former is significant only at the 5 per cent level, while the latter is significant at the 1 per cent level. The reason for this apparent discrepancy lies in the number of degrees of freedom for the two tests. These results show that there is some reason to doubt that the speed capabilities of the two aircraft types are equal, and also strong reason to doubt that individual planes within the same aircraft type are homogeneous as to speed capability. It should be stated that these data are entirely fictitious, and are intended only as an illustration; no conclusions about actual planes should be drawn from them.

If an experimental model involves one or more nesting factors and also one or more factors used in all combinations, it is called a partially hierarchal model. If the above experiment is repeated at high altitude, and the data for the two altitudes are considered together, the resulting model is partially hierarchal, with planes nesting within aircraft types but with altitudes occurring in all combinations with the other factors; that is, each plane of each aircraft type is tried at each altitude. The data in Table 8.4 represent the maximum true air speed on three passes of two planes each of two aircraft types (in horizontal flight) at high altitude.

In order to facilitate the calculation of the sums of squares, the data of Tables 8.1 and 8.4 are summarized in Table 8.5, where each entry represents the sum of three maximum true air speeds for a given plane at a given altitude.

Table 8.4

Aircraft Type 1		Aircraft Type 2	
Plane A	Plane B	Plane C	Plane D
480	480	570	600
490	440	580	580
470	460	560	590
1440	1380	1710	1770
2820		3480	

Table 8.5

	Aircraft Type 1			Aircraft Type 2			
Altitude	Plane A	Plane B	Sum	Plane C	Plane D	Sum	Total
Low	1350	1230	2580	1620	1740	3360	5940
High	1440	1380	2820	1710	1770	3480	6300
Total	2790	2610	5400	3330	3510	6840	12240

The total sum of squares (short-cut computation) is given by $(450)^2 + \cdots + (580)^2 + (480)^2 + \cdots + (590)^2 - (12240)^2/24 = 101,600.$

The sum of squares for aircraft types (T) is

$$[(5400)^2 + (6840)^2]/12 - (12240)^2/24 = 86,400.$$

The sum of squares for altitudes (A) is

$$[(5940)^2 + (6300)^2]/12 - (12240)^2/24 = 5,400.$$

The sum of squares for these two factors and their interaction is

$$[(2580)^2 + (3360)^2 + (2820)^2 + (3480)^2]/6 - (12240)^2/24 = 92,400.$$

Hence the sum of squares for the interaction (T x A) is

$$92,400 - 86,400 - 5,400 = 600.$$

The sum of squares for planes within aircraft type 1 $(P \mid T_1)$ is

$$[(2790)^2 + (2610)^2]/6 - (5400)^2/12 = 2,700.$$

The sum of squares for planes within aircraft type 2 $(P|T_2)$ is

$$[(3330)^2 + (3510)^2]/6 - (6840)^2/12 = 2,700.$$

Hence the sum of squares for planes within aircraft types $(P|T)$ is

$$2,700 + 2,700 = 5,400.$$

The sum of squares for altitudes in the entire experiment has already been found. Since altitudes do not nest within aircraft types, the sums of squares for altitudes within the two aircraft types will not figure in the final analysis. These sums of squares must, however, be computed as an intermediate step in finding the sum of squares for the interaction between planes and altitudes within aircraft types $(P \times A|T)$.

The sum of squares for altitudes within aircraft type 1$(A|T_1)$ is

$$[(2580)^2 + (2820)^2]/6 - (5400)^2/12 = 4,800.$$

The sum of squares for altitudes within aircraft type 2 $(A|T_2)$ is

$$[(3360)^2 + 3480)^2]/6 - (6840)^2/12 = 1,200.$$

The reader should note that the sum of these two sums of squares is not equal to the sum of squares for altitudes (A), but rather to the sum of the sums of squares for A and T x A.

The sum of squares for planes, altitudes and their interaction within aircraft type 1 is

$$[(1350)^2 + (1230)^2 + (1440)^2 + (1380)^2]/3 - (5400)^2/12 = 7,800.$$

The sum of squares for planes, altitudes and their interaction within aircraft type 2 is

$$[(1620)^2 + (1740)^2 + (1710)^2 + (1770)^2]/3 - (6840)^2/12 = 4,200.$$

Then the sum of squares for the interaction $(P \times A|T_1)$ is

$$7,800 - 2,700 - 4,800 = 300,$$

and the sum of squares for the interaction $(PxA|T_2)$ is

4,200 - 2,700 - 1,200 = 300.

Hence the sum of squares for the interaction (PxA|T) is

300 + 300 = 600.

The sum of squares for error is found by subtraction to be

101,600 - 86,400 - 5,400 - 600 -5,400 - 600 = 3,200.

Since aircraft types and altitudes each occur at two levels, the number of degrees of freedom for each is 2-1 = 1, and the number of degrees of freedom for their interaction is (2-1)(2-1) = 1. As in the previous example, the number of degrees of freedom for planes within aircraft types is 2(2-1) = 2. Since there are two altitudes, the number of degrees of freedom for interaction between planes and altitudes within aircraft types is 2(2-1)(2-1) = 2. Since there are three observations for each of four planes at each of two altitudes, the number of degrees of freedom for error is (2)(4)(3-1) = 16. Since there are 24 observations in all, the total number of degrees of freedom is 24-1 = 23.

The sums of squares, degrees of freedom, and mean squares for the various effects are given in Table 8.6.

Table 8.6

Source of Variation	SS	DF	MS
Aircraft types (T)	86,400	2-1 = 1	86,400/1 = 86,400
Altitudes (A)	5,400	2-1 = 1	5,400/1 = 5,400
Interaction (TxA)	600	(2-1)(2-1) = 1	600/1 = 600
Planes within types (P|T)	5,400	2 (2-1) = 2	5,400/2 = 2,700
Interaction (PxA|T)	600	2(2-1)(2-1) = 2	600/2 = 300
Error	3,200	2(4)(3-1) = 16	3,200/16 = 200
Total	101,600	24-1 = 23	

In making tests of significance, T is tested by the nesting factor P|T as in the preceding example, the fixed factor A is tested by PxA|T (its interaction with the random factor P|T), and the interaction TxA is tested by the nesting interaction PxA|T. The nesting factor P|T and its interaction PxA|T with the fixed factor A are both tested by error.

The tests are shown in Table 8.7.

Table 8.7

Source	F(= Variance Ratio)	$F_{.05}$	$F_{.01}$
T	86,400/2,700 = 32.00	18.51	98.49
A	5,400/ 300 = 18.00	18.51	98.49
TxA	600/ 300 = 2.00	18.51	98.49
P|T	2,700/ 200 = 13.50	3.63	6.23
PxA|T	300/ 200 = 1.50	3.63	6.23

These results confirm (and strengthen somewhat), the conclusions as to the significance of difference between aircraft types and between planes within types reached in the analysis of the preceding example. The F-ratio for altitudes falls just short of significance at the 5 per cent level, so there is some (but hardly conclusive) evidence that maximum true air speed depends on altitude. Neither of the interaction F-ratios is even close to being significant; hence we have no evidence that either of these interactions actually exists.

9. POOLING

Upon occasion a test can be made more sensitive through pooling of two (or more) mean squares preliminary to forming the F-ratio. This is possible because of the greater number of degrees of freedom for the pooled mean square, and consequently of a smaller F required for significance. Thus, pooling may enable the experimenter to obtain a significant difference which would not have been possible otherwise.

While the purpose of pooling is to obtain a more sensitive F-test by using a more accurate estimate of variation, it is conceivable that pooling could lead to the opposite result, a less sensitive test. With proper precaution, however, this will rarely happen. Also, improper and indiscriminate use of pooling may easily result in erroneous results, such as concluding that a factor is contributing significantly to the variation when it is not. Pooling of mean squares is generally useful if there is good reason to believe that they can be assumed to represent the same variation. If preliminary precaution is taken that this requirement is not seriously violated, the results based on pooled tests will usually be valid.

Section 2 discussed the decomposition of variation into parts and the degrees of freedom associated with them. Pooling, in contrast, is concerned with the synthesis of those parts of the variation which have been assumed to have arisen from the same source. This impliesa corresponding addition of the degrees of freedom associated with them.

Consider the following example of an Analysis of Variance table representing a factorial experiment where factor A is fixed (the population consisting of three levels), factor B consists of a random sample of four levels from an infinite population, and the number of replications is two.

Table 9. 1

Source of Variation	SS	DF	MS
A	24	3-1 = 2	24/2 = 12
B	27	4-1 = 3	27/3 = 9
Interaction (AxB)	12	(3-1)(4-1) = 6	12/6 = 2
Error	36	3(4)(2-1) = 12	36/12 = 3
Total	99	24-1 = 23	

In making tests of significance A is tested by error, B is tested by interaction (AxB), and AxB is tested by error. The results are given in Table 9. 2.

Table 9. 2

Source	F (= Variance Ratio)	$F_{.05}$	$F_{.01}$
A	12/3 = 4.00	3. 88	6. 93
B	9/2 = 4. 50	4. 76	9. 78
AxB	2/3 = 0. 67	3. 00	4. 82

It is noticed here that the F-ratio for A is significant at the 5 per cent level, and the F-ratios for B and AxB are not significant.

If there is good reason to believe that the interaction has no effect on the variation, then the interaction mean square and the error mean square can be pooled (since they would then be estimating the same variation, namely that due to chance only) so as to afford a more sensitive test of A. Since the interaction mean square is, before pooling, the appropriate denominator term of the F-ratio for A, the other mean square, error, which is being considered for pooling with it will be referred to as the doubtful mean square. As corroborating evidence that the interaction has no contribution (other than chance variations), the F-ratio of interaction to error must be not significant at the 5 per cent level. It must further satisfy the more restrictive requirement that it be not significant at an even lower level of significance, say 25 per cent or 50 per cent, before pooling is justified. This more restrictive test to determine feasibility of pooling is usually referred to as a preliminary test of significance. What level of significance one should use for the preliminary test has been the subject for recent theoretical research. This is discussed under the two cases given below.

CASE I: The expected value of the denominator mean square for the factor to be tested is greater than or equal to that of the doubtful mean square.

For case I, the following recommendations (stated according to two alternative conditions) concerning when to pool are based on the results given in WADC Technical Report 55-244, Volume I. Their validity has been demonstrated only for certain two-factor random and mixed models.

Condition (a) - The number of degrees of freedom for the factor to be tested is greater than or equal to that for the denominator term in its F-ratio. The number of degrees of freedom for the doubtful mean square is greater than or equal to five times that for the denominator term (of the factor to be tested).

Condition (b)* - Either the number of degrees of freedom for the factor to be tested is less than that for the denominator term in its F-ratio, or the number of degrees of freedom for the doubtful mean square is less than five times that for the denominator term (of the factor to be tested).

Under condition (a), pooling is usually justified if the F-ratio of the denominator mean square (for the factor to be tested) to the doubtful mean square is not significant at the 50 per cent level. Under condition (b), pooling is usually justified if the F-ratio of the denominator mean square (for the factor to be tested) to the doubtful mean square is not significant at the 25 per cent level.

Consider again the example described above. The expected value of the interaction mean square is always greater than or equal to that

*Of the two conditions, condition (b) is the one more commonly satisfied.

of the error mean square (which, in the example, is the doubtful mean square.) Thus, the above example comes under case I. Also the number of degrees of freedom for B, 3, is less than that for interaction, 6. Hence, condition (b) holds. The F-ratio for interaction (with respect to error) is

$$F = \frac{2}{3} = 0.67,$$

which is not significant at the 25 per cent level ($F_{.25} = 1.53$). On the basis of this evidence a pooled mean square of the interaction and error is desirable for testing the significance of factor B. The pooled mean square is simply the weighted average of the two with corresponding degrees of freedom for weights.

Thus, the pooled mean square to test B is

$$\frac{(6x2)+(12x3)}{6+12} = \frac{48}{18} = 2.67,$$

with 18 degrees of freedom. The altered F-ratio for B now is

$$F = \frac{9}{267} = 3.37$$

which is now significant at the 5 per cent level ($F_{.05} = 3.16$), thus showing a significant difference for the levels of factor B. This result could not be arrived at without pooling.

Consider the example of section 6, except that the columns factor is taken as random instead of fixed. The mean squares (see Table 6.2) again satisfy condition (b). As stated in section 7 the proper denominator term for testing rows is interaction. Consider the possibility of pooling error with interaction for testing rows. Here the F-ratio for interaction (with respect to error) is significant at the 5 per cent level. Since the value of $F_{.25}$ is always less than the value of $F_{.05}$, this insures that the F-ratio is also significant at the 25 per cent level. Consequently, one does not pool here.

It is not known at the present time whether the above recommendations are valid for other models under case I. However, they appear to be quite reasonable with respect to ease of manipulation and to give satisfactory results. Until further research in this area yields more desirable methods, it is conjectured that the above procedure is not far wrong when applied to more general models under case I.

As a final example for models under case I consider that of section 8. The F-ratio for altitude (A) was not significant at the 5 per cent level when tested by $PxA|T$; neither was the F-ratio for interaction TxA when tested by $PxA|T$ (see Table 8.7). The question naturally arises as to whether pooling $PxA|T$ with error would increase the sensitivity of these two tests. Pooling would only be valid if it is reasonable to assume that the $PxA|T$ interaction does not contribute to the variation (except chance variation represented by error). Here the expected value of the $PxA|T$ mean square is greater than or equal to that of the error mean square (case I).

This example is of a more general model than those for which the recommendations have been shown to hold. However, pooling would definitely be helpful here. Thus, a preliminary test is called for. In the absence of information concerning the proper level of significance for the preliminary test, one chooses 25 per cent or 50 per cent according to the stated recommendations (even though, strictly speaking, they may not be the best to use.)

Both the tests for A and for TxA satisfy condition (b). Therefore, 25 per cent is the level of significance for the preliminary test of the F-ratio of $PxA|T$ to error. Inasmuch as

$$F = \frac{300}{200} = 1.50$$

is not significant at the 25 per cent level ($F_{.25} = 1.51$), pooling is probably valid. The pooled mean square of $PxA|T$ and error is

$$\frac{2x300 + (16x200)}{2+16} = \frac{3800}{18} = 211.$$

The altered F-ratio for A is

$$F = \frac{5400}{211} = 25.6,$$

which is now significant at the 1 per cent level ($F_{.05} = 4.41$, $F_{.01} = 8.28$).

Using a pooled test thus gives a significant result and one can now conclude much more definitely than before that altitude affects the maximum airspeed. The altered F-ratio for TxA is

$$F = \frac{600}{211} = 2.84 ,$$

which still is not significant at the 5 per cent level ($F_{.05} = 4.41$). Thus, while pooling here has made the test more sensitive ($F_{.05} = 4.41$ as contrasted with the pre-pooling value, $F_{.05} = 18.51$), one still cannot conclude that the TxA interaction contributes significantly to the variation.

CASE II: The expected value of the denominator mean square for the factor to be tested is greater than or equal to that of the doubtful mean square.

Here the preliminary test involves the F-ratio which is the reciprocal of that in case I, i.e. the F-ratio of the doubtful mean square to the denominator mean square of the factor to be tested. For this case, there is no evidence at this time as to what a satisfactory level of significance for the preliminary test should be. One possibility is to use the 25 per cent level or the 50 per cent level according as condition (b) or condition (a) holds. It may or may not be a reasonably good procedure to follow.

10. FRACTIONALLY REPLICATED EXPERIMENTS

An experiment in which at least one observation is made for each possible combination of levels of the factors is said to be completely replicated. An experiment in which observations are made for only a part of the possible combinations of levels of the factors is said to be fractionally replicated. A by-product of fractional replication is the inability to separate the effects of certain main factors and interactions. Two or more effects which cannot be separated are said to be confounded, and each is said to be an alias of the other(s).

In many cases where the unit cost of observations is high and the number of possible combinations of levels of the factors is large, the total cost of complete replication is prohibitive. In such cases, a part (but not all) of the information which would be obtained from a completly replicated experiment can be obtained from a fractionally replicated experiment at a much lower total cost. This is true especially when it is known (or can safely be assumed) that the higher-order interactions (interactions of several factors) do not exist. Then the main effects (and sometimes also the low-order interactions) can be tested, even though they are confounded with higher-order interactions (known or assumed to be non-existent).

Consider, for example, an experiment designed to determine the effect of four alloying elements on the tensile strength of titanium. Each alloying element will be studied at two levels (none and some standard percentage). Since this experiment will involve four factors each at two levels, which could be studied in all combinations, it will be spoken of as a 2^4 factorial experiment. Complete replication would require $2^4 = 16$ observations, and would result in the analysis shown in Table 10. 1, where A, B, C and D are the alloying elements.

Table 10. 1

Source of Variation	DF
Main effects (A, B, C, D)	4
Two-factor interactions, (AB, AC, AD, BC, BD, CD)	6
Three-factor interactions (ABC, ABD, ACD, BCD)	4
Four-factor interaction (ABCD)	1
Total	$2^4 - 1 =$ 15

Even with complete replication (one observation for each combination of levels of the factors), one would have to make some assumption about higher-order interactions in order to be able to perform any significance tests. One might, for example, assume that the three- and four-factor interactions really do not exist, so that the mean squares found for them are really estimates of experimental error. In this case the main effects and two-factor interactions (each with 1 degree of freedom) would be tested by the error mean square (with 5 degrees of freedom) obtained by pooling the three-and four-factor interactions.

If one can assume also that the two-factor interactions are non-existent, information about the main effects in the experiment described in the preceding paragraph can be obtained from a one-half replicate of the experiment, requiring only (1/2)(16) = 8 observations. In such an experiment, the mean (designated by I) will be confounded with the four-factor interaction. This fact will be represented symbolically by the confounding relation I = ABCD.

An effect will be denoted by a capital letter, while the presence of an alloying element in an observed specimen will be denoted by the corresponding lower case letter. Thus A denotes the effect of alloying element A, while a denotes that A is present as an alloying element in the specimen under observation. The symbol (1) will signify that none of the alloying elements is present. The observations which must be made in a half-replicate experiment are all those involving an odd number (or alternatively, an even number) of lower case letters. Thus for a half-replicate of a 2^4 factorial experiment, the observations must be either

a, b, c, d, abc, abd, acd, bcd

or

(1), ab, ac, ad, bc, bd, cd, abcd.

The former set of observations will be used in an example, for which the data are given in Table 10.2.

Table 10.2

a	b	c	d	abc	abd	acd	bcd	Total
16	24	18	14	37	39	33	35	216

In this experiment, the main effect of each factor is confounded with the interaction of the other three, and the interaction of each pair of factors is confounded with the interaction of the other two. If it is known or can be assumed that the interactions are all non-existent, then one can test the main effect of each factor (with 1 degree of freedom) by comparison of its mean square with the mean square for error (with 3 degrees of freedom) obtained by pooling the two-factor interactions.

The overall effect of each factor can be found by subtracting the sum of all observations not involving that factor from the sum of all observations which do involve that factor, thus:

Effect of A = (16 + 37 + 39 + 33) - (24 + 18 + 14 + 35) = 125-91 = 34,
Effect of B = (24 + 37 + 39 + 35) - (16 + 18 + 14 + 33) = 135-81 = 54,

Effect of C = (18 + 37 + 33 + 35) - (16 + 24 + 14 + 39) = 123 - 93 = 30,
Effect of D = (14 + 39 + 33 + 35) - (16 + 24 + 18 + 37) = 121 - 95 = 26.

The sum of squares for each effect may be found by dividing the square of the overall effect by the total number of observations, thus:

$$\text{SS for A} = (34)^2/8 = 144.5; \quad \text{SS for B} = (54)^2/8 = 364.5;$$
$$\text{SS for C} = (30)^2/8 = 112.5; \quad \text{SS for D} = (26)^2/8 = 84.5.$$

The total sum of squares for the entire set of data is $(16)^2 + (24)^2 + (18)^2 + (14)^2 + (37)^2 + (39)^2 + (33)^2 + (35)^2 - (216)^2/8 = 724$.

The sum of squares for the three pairs of two-factor interactions (to be interpreted as error) is found by subtraction to be

$$724 - 144.5 - 364.5 - 112.5 - 84.5 = 18.$$

The analysis of variance for the data of Table 10.2 is shown in Table 10.3

Table 10.3

Source	SS	DF	MS	F (Variance Ratio)
A	144.5	1	144.5/1 = 144.5	144.5/6 = 24.08
B	364.5	1	364.5/1 = 364.5	364.5/6 = 60.75
C	112.5	1	112.5/1 = 112.5	112.5/6 = 18.75
D	84.5	1	84.5/1 = 84.5	84.5/6 = 14.08
Error	18	3	18 /3 = 6	
Total	724	7		

For each F-ratio there are 1 and 3 degrees of freedom, so that $F_{.05} = 10.13$ and $F_{.01} = 34.12$. Thus the effect of factor B is significant at the 1 per cent level, while the effects of the other factors are significant at the 5 per cent level, but not at the 1 per cent level. The overall effect of each factor is positive; that is, the presence of each alloying element tends to increase the tensile strength.

The example used here was chosen for simplicity; a recent Air Force problem (see WADC TN 55-14) involved a one-eighth replicate of a 2^{10} factorial experiment.

11. RANDOMIZATION AND THE STANDARD STATISTICAL DESIGNS.

An experimental design is a blueprint according to which an experiment is patterned. A statistical design is an experimental model which includes a further essential feature--a randomization process. It furnishes the basis upon which appropriate statistical tests and inferences can be made. The order in which observations are made, the method of selection of experimental material, the factors to be considered and their levels, the number of experimental units available, the environmental influences, the characteristics of the experimental material, the particular randomization procedure to be used in assigning experimental units to the conditions under study, and the process of recording data are all considerations to be taken into account when evolving an appropriate statistical design for a given experimental situation. Since the design is supposedly representative of the actual experiment, certain assumptions made in the design must be approximated to a satisfactory degree by the experiment actually performed. Otherwise, the inferences made would be invalid. The statistical design is then inappropriate for the experiment, or the experiment is inappropriate for the design.

Some commonly known statistical designs of a simple nature are described in this section. From past experience these have been shown to be very useful. Historically, they were developed for agricultural experimentation. However, their use has expanded into other natural sciences and they are becoming increasingly useful for experimentation in such fields as engineering, chemistry, astronomy, psychology, and other physical and biological sciences, both pure and applied. Four typical designs will be considered and their advantages and disadvantages discussed:

(1) completely randomized design,
(2) randomized block design,
(3) Latin square design,
(4) factorial design.

The completely randomized design is essentially a one-factor model with replication. As indicated by the name the assignment of an experimental unit to a given level of the factor is completely random in the statistical sense. Random (in the statistical sense) means that any one of the experimental units has an equal chance of being chosen for a given level of the factor. Table 3. 1 is an example of a completely randomized design, if it is considered that the twelve items were made from twelve pieces of raw material (experimental units) obtained from a common source, each factory being assigned four pieces at random.

The following example is taken from Freeman.* For purposes of comparison with the other designs to be discussed these same data will be used over again. Of course, for each design the context will be different as will be the experimental situation and the assumptions made; only the numerical values are kept the same so as to provide suitable contrast among the different designs.

An experiment was conducted to compare the effects of five different types of grids, A, B, C, D, E, on the vacuum of radio tubes. A total of twenty-five tubes was used for the experiment.

Suppose for the present that one has no knowledge whatever as to what causes other than grid type might affect the observations. The only safeguard in such a situation is to make sure that these do not bias the results. And the only effective way of doing this is to randomize. A simple procedure for randomizing is:

(1) Assign numbers 1, 2, ···, 25 to the twenty-five tubes.

(2) Put these numbers on slips of paper, throw them into a hat, and shuffle well. Pick out the slips of paper. Assign the first five to grid A, the second five to grid B, and so on.

The drawbacks to such a procedure are that because of non-uniformity of size of paper slips, and of the difficulty of effecting thorough shuffling, each slip may not have an equal chance of being chosen for a given grid, and true randomness may not be attained. A much more dependable randomization procedure is to use a table of random numbers.* One possibility of using such a table for this procedure is given below.

(1) Using two-digit random numbers, assign random numbers 00-03 to tube 1, numbers 04-07 to tube 2,···, 96-99 to tube 25.

(2) Starting anywhere in the table of random numbers and proceeding in a systematic order, pick out the tubes according to the

*Freeman, H. A., Industrial Statistics, p. 52.

*E. g., Tippett, L. H. C., Random Sampling Numbers.

two-digit numbers chosen, ignoring any number if it corresponds to a tube already chosen. Assign the first five tubes to grid A, the second five to grid B, and so on.

The observations are then taken according to the above plan. The data as given in terms of a relative measure of vacuum (obtained by subtracting 90.0 from each of Freeman's data) may be as in Table 11.1.

Table 11.1

	A	B	C	D	E
	3.6	5.3	4.5	6.8	4.6
	5.3	6.9	7.0	8.2	7.8
	7.0	5.8	7.8	7.2	8.0
	3.7	7.3	7.0	7.2	5.0
	8.0	7.7	8.3	7.9	8.9
Sum	27.6	33.0	34.6	37.3	34.3
Mean	5.52	6.60	6.92	7.46	6.86

The analysis of variance for this completely randomized experiment is given below in Table 11.2.

Table 11.2

Source of Variation	SS	DF	MS	F	$F_{.05}$
Difference among grids	10.25	4	2.56	1.16	2.87
Tubes within grids	44.18	20	2.21		
Total	54.43	24			

The difference among grids is not significant at the 5 per cent level.

The analysis of variance for the general completely randomized experiment with k levels for the factor under consideration and n replications is given by Table 11.3.

Table 11.3

Source of Variation	SS	DF	MS	F
A = Difference among k levels	$n\sum_i (x_{i.} - x_{..})^2$	k-1	SS_A/DF_A	MS_A/MS_E
E = Error	$\sum_i \sum_j (x_{ij} - x_{i.})^2$	k(n-1)	SS_E/DF_E	
Total	$\sum_i \sum_j (x_{ij} - x_{..})^2$	nk-1		

Randomization insures that under repeated observations a treatment will not be continually favored or handicapped by a source of variation which the experimenter had not taken account of. Randomization is also important from another point of view. It is the only part of the experimental procedure which makes use of the laws of chance. Since the laws of chance in turn are associated with the frequency distribution on which the statistical tests of significance are based, randomization provides the necessary foundation for a valid test of significance. This is exactly the reason why a systematic arrangement fails. Thus, randomization is absolutely necessary to insure valid estimates of variation.

While randomization has made significance tests possible, the completely randomized design nevertheless represents a very "insensitive" experiment, because no attempt has been made to decrease the error variance, i.e., to increase the precision. The precision can be maximized by separating out the variation of other factors suspected of contributing significantly to the variation.

One type of experimental procedure used to improve the precision or sensitivity of an experiment is to divide the group of experimental units into subgroups which are homogeneous within themselves. This is possible only if one is able to recognize that there is homogeneity with respect to some criterion. Because of a restriction in the randomization procedure that will be described later in this section, the number of experimental units per subgroup must be equal to the number of levels of the factor being studied (or to the total number of combined levels, in the case where more than one factor is being studied). All levels of the factor are then assigned, by a random procedure such as that described above, to each subgroup of experimental units. These subgroups are referred to as blocks, and the design is known as a randomized block design.*

In a randomized block design the variation due to differences among blocks (block effect) is separated out from the error variation. The characteristics of such a design are:

(1) blocks are as different as possible;

(2) experimental units within blocks are as similar as possible.

*For such a design to be efficient, the factor x block interaction must be negligible. This point should be taken into account if one desires to use the randomized block design.

Consider now the example described on the effect of different grid types on the vacuum of radio tubes. In order to increase the precision one should attempt to remove other causes of variation which may greatly influence the vacuum. For example, the machines which seal the tubes may vary in their ability to produce a vacuum, the operators that handle the machines may differ in comparative skill, and temperature and humidity may be important factors.

One method often advocated for removing such causes is to have one operator use one machine for sealing all types, under standard temperature and humidity conditions. (This is known as standardization of experimental techniques). However, as will be discussed more fully later, the limitations of such an approach are severe. For example, it may take too long to use just one machine and one operator. Besides, one is often vitally interested in answering other questions, such as "do differences in operator skill significantly affect the vacuum?"

Another equally valid method which provides an answer to this question is to assign (by a randomization procedure) Type A, B, C, D, E grids on 5 tubes which will be sealed by Operator 1, Type A,···, E on 5 tubes to be sealed by Operator 2,···, Type A,···, E on 5 tubes to be sealed by Operator 5*. This is a randomized block design with each operator representing a block. Since a grid x operator interaction is not a likely prospect, this design would probably be efficient for the example.

One possible randomization procedure here is:

(1) Choose 5 tubes for Operator 1, 5 tubes for Operator 2, ···, 5 tubes for Operator 5 by the method described on bottom of page 30 (replacing "grids" by "operators" in that description).

(2) Assign the 5 tubes for Operator 1 at random to grids A, B, C, D, E, using a table of random numbers; repeat for Operator 2 through Operator 5.

*In a randomized block design it is not necessary that the number of blocks equal the number of levels of the factor being studied, though this is the case for this example. Here, the number of operators (blocks) = the number of grid types = 5. On the other hand, it is necessary that the number of experimental units per block be equal to the number of levels of the factor being studied.

The principle involved here is randomization subject to one restriction; namely, that a given level of the factor being studied occurs once and only once in a given block.

Steps (1) and (2) can be combined into an even simpler randomization procedure:

(1) Associate code number-letters 1A, 1B, ···, 5E with random-number intervals 00-03, 04-07, ···, 96-99, where the code number refers to the operator, and the code letter refers to the type of grid. Taking the tubes in any order, assign the first tube to the appropriate operator and grid according to the first random number chosen (use a table of two-digit random numbers;throw out all repetitions); and so on.

Following the above randomization procedure, the observations are taken, and the resulting set of data may be given in Table 11.4.

Table 11.4

	Grid						
Operator	A	B	C	D	E	Sum	Mean
1	7.0	5.8	7.8	7.2	8.0	35.8	7.16
2	8.0	7.7	8.3	7.9	8.9	40.8	8.16
3	3.6	5.3	4.5	6.8	4.6	24.8	4.96
4	3.7	7.3	7.0	7.2	5.0	30.2	6.04
5	5.3	6.9	7.0	8.2	7.8	35.2	7.04
Sum	27.6	33.0	34.6	37.3	34.3	166.8	
Mean	5.52	6.60	6.92	7.46	6.86		6.67

The analysis of variance for this experiment is given in Table 11.5.

Table 11.5

Source of Variation	SS	DF	MS	F	$F_{.05}$	$F_{.01}$
Difference among grids	10.25	4	2.56	2.78	3.01	4.77
Difference among operators	29.59	4	7.40	8.04	3.01	4.77
Discrepance	14.69	16	.92			
Total	54.43	24				

Note that the difference among operators is significant at the 1 per cent level. Since this significant effect has been removed from the error variation, a much more sensitive test can be made for grid effect. However, the F-value for grids, while close to the 5 per cent value, is still not significant. Contrast Table 11.5 with Table 11.2.

The analysis of variance table for the general randomized block experiment with k levels for the factor under consideration and n blocks containing k experimental units is given by Table 11.6.

Table 11.6

Source of Variation	SS	DF	MS	F
A = Difference among k levels	$n\sum_i (x_{i.} - x_{..})^2$	k-1	SS_A/DF_A	MS_A/MS_D
B = Blocks	$k\sum_j (x_{.j} - x_{..})^2$	n-1	SS_B/DF_B	MS_B/MS_D
D = Discrepance	$\sum_i \sum_j (x_{ij} - x_{i.} - x_{.j} - x_{..})^2$	(k-1)(n-1)	SS_D/DF_D	
Total	$\sum_i \sum_j (x_{ij} - x_{..})^2$	kn-1		

The advantage of the randomized block design is the elimination of the effect on the error mean square of large differences among experimental units. The error sum of squares is diminished by the amount separated out for the block effect, and a smaller estimate of the remaining variation (discrepance) results. The disadvantage of using a randomized block design occurs when the factor x block interaction is not negligible.

Consider now a third classical design, the Latin square. Whereas the randomized block design eliminates the effect of one environmental factor other than the factor being studied, the Latin square eliminates two. For convenience, designate these two factors by "row" and "column". Here a double restriction is imposed on the experimental procedure: a given level of the factor being studied occurs once and only once within a row and once and only once within a column. This restriction implies that the number of rows and the number of columns are each equal to the number of levels of the factor being studied.

In the example the Latin square design would enable one to eliminate differences among machines (column effect) in addition to eliminating operator differences (row effect). As an illustration of such a design one now considers five operators and five machines in connection with the five grids. In order for the Latin square to be efficient, it must be reasonable to assume that none of the possible interactions is significant. Otherwise the design is inadequate.

A possible randomization procedure is the following:

(1) For Operator 1, assign the five grids to the five machines at random;

(2) For Operator 2, assign the five grids to the five machines at random except that no grid is allowed to occur which has already been assigned to Operator 1 on the same machine (if such an event occurs, throw out such a random number);

(3) Continue in the same manner, employing the double restriction that a grid can occur only once in a given row and only once in a given column, until all rows and columns are filled.

Following such a randomization procedure, the observations are taken, and the resulting set of datá may be given in Table 11.7.

Table 11.7

Operator	Machine				
1	8.0(E)*	5.8(B)	7.2(D)	7.0(A)	7.8(C)
2	8.3(C)	7.9(D)	7.7(B)	8.9(E)	8.0(A)
3	3.6(A)	4.5(C)	4.6(E)	5.3(B)	6.8(D)
4	7.2(D)	5.0(E)	3.7(A)	7.0(C)	7.3(B)
5	6.9(B)	5.3(A)	7.0(C)	8.2(D)	7.8(E)

*Grid type

The analysis of variance for this experiment is given by Table 11.8.

Table 11. 8

Source of Variation	SS	DF	MS	F	$F_{.05}$	$F_{.01}$
Difference among grids	10.25	4	2.56	13.47	3.26	5.41
Difference among operators	29.49	4	7.40	38.94	3.26	5.41
Difference among machines	12.42	4	3.11	16.37	3.26	5.41
Discrepance	2.27	12	.19			
Total	54.53	24				

For this experiment, it is found that grid effect, operator effect, and machine effect are all significant at the 1 per cent level. Contrast Table 11. 8 with Tables 11. 2 and 11. 5. Thus, for this example, the Latin square arrangement permitted the statistical demonstration that differences among grids had a significant effect on vacuum, a result which could not be determined by either the randomized block design or the completely randomized design. It should not be tacitly assumed, however, that the Latin square design is always to be preferred to the other two. As mentioned at the beginning of this section the appropriateness of a given design depends on many considerations, including the restrictions necessarily imposed by its use. Notice that the restrictions on the nature of the observations become increasingly stringent as one progresses from the completely randomized design, through the randomized block design, to the Latin square design. Thus because of the increased restrictions, the Latin square design may not be the appropriate one at all to use for a given experimental situation.

The analysis of variance table for the general Latin square arrangement with k levels for the factor under consideration, k rows, and k columns is given by Table 11. 9.

For greater precision, the Latin square can be replicated.

The Latin square design is especially suited to study the variation due to 4 to 8 levels of the factor under consideration, i. e. , a small selected number, where the effect of varying levels of two other factors needs to be considered simultaneously. As in the case of randomized blocks, the disadvantage in using a Latin square design occurs when any of the interactions are not negligible, since in such a case the design is an inefficient one. In the event of non-negligible interaction(s) , or when it is desired to study a factor under varying conditions of a large number of other factors, a factorial design is a more appropriate one to use. The factorial design involves

Table 11.9

Source of Variation	SS	DF	MS	F
A = Difference among k levels	$k\sum_{t} (x_t - x_{..})^2$[*]	k-1	$SS_A/k-1$	MS_A/MS_D
B = Difference among rows	$k\sum_{i} (x_{i.} - x_{..})^2$	k-1	$SS_B/k-1$	MS_B/MS_D
C = Difference among columns	$k\sum_{j} (x_{.j} - x_{..})^2$	k-1	$SS_C/k-1$	MS_C/MS_D
D = Discrepance	$\sum_{i}\sum_{j}(x_{ij} - x_{..})$ $- SS_A - SS_B - SS_C$	(k-1)(k-2)	$SS_D/(k-1)(k-2)$	
Total	$\sum_{i}\sum_{j}(x_{ij} - x_{..})^2$	k^2-1		

observations on all possible combinations of the levels of all factors being considered. For example, a three-factor factorial arrangement with two replications and 2 levels of factor A, 4 levels of factor B, 3 levels of factor C, involves 2x2x4x3 = 48 observations.

The data of Table 11.4, representing a one-factor randomized block design, illustrate essentially a two-factor factorial arrangement which is unreplicated. The data given in Table 6.1 and the corresponding analysis of variance Tables 6.2, 6.3 illustrate a two-factor factorial arrangement with replication. Note that Table 6.3 indicates a significant interaction at the 5 per cent level. The analysis of variance table for a three-factor factorial design with a levels for factor A, b levels for factor B, c levels for factor C, and n replications, is given in Table 11.10. No F-ratios are indicated since they will depend on whether the factors are fixed or random.

There are abc possible combinations of levels for this three-factor experiment. The randomization procedure may be carried out in the following manner: choose at random out of abcn experimental units,

[*] x_t is the average of the k values pertaining to the tth level of the factor being studied.

Table 11.10

Source of Variation	SS	DF	MS*
Single Factors:			
A	$nbc \sum_{i} (x_{i\cdot\cdot\cdot} - x_{\cdot\cdot\cdot\cdot})^2$	a-1	
B	$nac \sum_{j} (x_{\cdot j\cdot\cdot} - x_{\cdot\cdot\cdot\cdot})^2$	b-1	
C	$nab \sum_{k} (x_{\cdot\cdot k\cdot} - x_{\cdot\cdot\cdot\cdot})^2$	c-1	
Interaction between 2 factors:			
AxB	$nc \sum_{i} \sum_{j} (x_{ij\cdot\cdot} - x_{i\cdot\cdot\cdot} - x_{\cdot j\cdot\cdot} + x_{\cdot\cdot\cdot\cdot})^2$	(a-1)(b-1)	
AxC	$nb \sum_{i} \sum_{k} (x_{i\cdot k\cdot} - x_{i\cdot\cdot\cdot} - x_{\cdot\cdot k\cdot} + x_{\cdot\cdot\cdot\cdot})^2$	(a-1)(c-1)	
BxC	$na \sum_{j} \sum_{k} (x_{\cdot jk\cdot} - x_{\cdot j\cdot\cdot} - x_{\cdot\cdot k\cdot} + x_{\cdot\cdot\cdot\cdot})^2$	(b-1)(c-1)	
Interaction among all three factors:			
AxBxC	$n \sum_{i} \sum_{j} \sum_{k} (x_{ijk\cdot} - x_{ij\cdot\cdot} - x_{i\cdot k\cdot} - x_{\cdot jk\cdot} + x_{i\cdot\cdot\cdot} + x_{\cdot j\cdot\cdot} + x_{\cdot\cdot k\cdot} - x_{\cdot\cdot\cdot\cdot})^2$	(a-1)(b-1)(c-1)	
Error:			
E	$\sum_{i} \sum_{j} \sum_{k} \sum_{\eta} (x_{ijk\eta} - x_{ijk\cdot})^2$	abc(n-1)	
Total	$\sum_{i} \sum_{j} \sum_{k} \sum_{\eta} (x_{ijk\eta} - x_{\cdot\cdot\cdot\cdot})^2$	abcn-1	

*MS = SS/DF.

the n experimental units that are to be given a certain combination; continue choosing until all abc combinations are assigned their n experimental units.

Many chemical and engineering experiments involve only variations of single factors with control of the other factors at a fixed level. As mentioned earlier in this section this is referred to as <u>standardization</u>

of experimental techniques. There are at least three important advantages of a factorial experiment over a standardized experiment:

(1) more efficiency,
(2) more information,
(3) broader inductive basis.

To illustrate these points consider the case of four factors. A factorial experiment measuresthe effect of each of the four with the same precision as if the entire experiment had been devoted to one factor only; thus only one fourth as many observations are used as for the standardized arrangement, and greater efficiency is achieved.

In addition, an evaluation can be made of all the interactions among these factors with the same precision, whereas with a series of experiments on each factor singly no such information can be obtained. The interactions may or may not contribute significantly but this information nevertheless needs to be known; thus more information is gained.

Any conclusion concerning a given factor based upon an experiment in which the other important factors are varied has a much broader inductive interpretation than that which is based upon experimentation in which these are kept standardized (regardless of the amount of experimentation involved). The factorial design deliberately varies these factors. Since a highly standardized experiment furnishes direct information only for the narrow range of conditions achieved by it, it weakens, rather than strengthens, the grounds for inferring a like result under varied conditions. This is why, in practice, a standardized experiment may not be the desirable design to use, though it is often strongly advocated because of its simplicity.

The factorial design needs to be replicated to insure an independent estimate of error. Often, in unreplicated experiments it is assumed that the highest order interaction is negligible to allow for suitable tests. Whether this assumption is a good one or not depends upon the particular experimental situation.

To conclude this section, a partial list of some more complex designs (which have found use under more specialized experimental requirements) is mentioned below:

(1) cross-over designs,
(2) Graeco-Latin squares (randomization under triple restrictions),
(3) split-plot designs,
(4) balanced incomplete block designs,
(5) lattice designs,
(6) Youden squares.

12. MULTIPLE COMPARISONS

The basic F-test in an analysis of variance determines whether there is a significant difference among a group of means, but it cannot tell which means differ significantly from which others. The latter is often what the investigator really wants to know. Various multiple comparisons tests have been proposed to determine whether each mean differs significantly from each other. Perhaps the best of these is the Newman-Keuls test, which will be described and illustrated in this section.

Suppose the F-test has shown that the means for the levels of a particular factor differ significantly at the 5 per cent level (or the 1 per cent level). If one desires to know which means differ significantly from which others, he should arrange the means in order from smallest to largest. Let m be the number of levels of the factor under consideration, let s^2 be the mean square used as the denominator of the F-ratio for testing that factor, let n_2 be the number of degrees of freedom for s^2, and let k be the number of observations at each level of the factor under consideration. Then the significance of p successive ordered means ($p=m, m-1, \cdots, 2$) can be tested by comparing their range (the difference between the largest and smallest of the p means) with the critical range for p means. To obtain the critical range of p out of m ordered means for the Newman-Keuls procedure, one multiplies (for the desired significance level) the studentized range* of p observations with n_2 degrees of freedom for s^2 by the standard error of the mean, $s_{\bar{x}} = \sqrt{s^2/k}$.

*The 5 per cent and 1 per cent levels of the studentized range have been tabulated in Pearson and Hartley, Biometrika Tables for Statisticians, Biometrika Office, London, 1954.

Usually the Newman-Keuls procedure will show the entire set of m means to be significant at the same level as does the F-test. In fact, the Newman-Keuls test on the entire set may be used as a substitute for the F-test. If there is a significant difference among the m means, then one proceeds to make the test on (m-1) means, omitting first the largest, then the smallest (or vice versa). If a significant difference is found, one then tests groups of (m-2) means, and so on until no further significant differences are found. Each group of means found not to differ significantly is underscored by a single line. Thus two means not underscored by the same line do differ significantly. It should be noted that p means are not declared to be significantly different if they all belong to a larger group of means found to be not significant, even though their range may exceed the critical range of p means.

As an example of the Newman-Keuls procedure, consider the data shown in Table 12.1, which might represent the survival time (in days) of rats on five diets.

Table 12.1

	A	B	C	D	E	
	1	6	7	11	6	
	2	4	5	11	8	
	3	8	3	8	7	
Total	6	18	15	30	21	90
Mean	2	6	5	10	7	6

An analysis of variance of these data yields the results shown in Table 12.2.

Table 12.2

Source of Variation	SS	DF	MS	F
Difference among diets	102	4	25.5	9.81
Rats within diets	26	10	2.6	
Total	128	14		

Since $F_{.01} = 5.99$ for 4 and 10 degrees of freedom, the difference among diets is significant at the 1 per cent level.

In applying the Newman-Keuls procedure, one first arranges the means in order from smallest to largest, thus:

A	C	B	E	D
2	5	6	7	10

Next one computes the standard error of the mean,

$$s_{\overline{x}} = \sqrt{2.6/3} = \sqrt{0.867} = 0.931$$

The 5 per cent and 1 per cent points for the studentized range of p observations (p = 2, 3, 4, 5) with 10 degrees of freedom for s^2 are given in Table 12.3.

Table 12.3

	p = 2	p = 3	p = 4	p = 5
5%	3.15	3.88	4.33	4.65
1%	4.48	5.27	5.77	6.14

The critical ranges, which are the products of $s_{\overline{x}} = 0.931$ and the corresponding values in Table 12.3, are given in Table 12.4.

Table 12.4

	p = 2	p = 3	p = 4	p = 5
5%	2.93	3.61	4.03	4.33
1%	4.17	4.91	5.37	5.72

The range of all five means (10 - 2 = 8) is significant at the 1 per cent level, since it is greater than 5.72. The range of four means, excluding D, (7 - 2 = 5) and the range of four means, excluding A, (10 - 5 = 5) are both significant at the 5 per cent level, but not at the 1 per cent level, since 5.37 > 5 > 4.03. The range of three means, excluding D and E, (6 - 2 = 4) and the range of three means, excluding A and C, (10 - 6 = 4) are significant at the 5 per cent level, but the range of three means, excluding A and D, (7 - 5 = 2) is not significant at the 5 per cent level, since 4 > 3.61 > 2. The range of two means, A and C, (5 - 2 = 3) and the range of two means, E and D, (10 - 7 = 3) are significant at the 5 per cent level, since 3 > 2.93. The ranges of two means, C and B, and of two means, B and E, are not considered for possible significance, since both of these pairs are subsets of a set of three means, excluding A and D, whose range was found to be not significant.

The results of the 1 per cent level test may be summarized as follows:

$$\begin{array}{ccccc} A & C & B & E & D \\ 2 & 5 & 6 & 7 & 10 \\ \cline{1-4} \\ \cline{2-5} \end{array}$$

A similar summary for the 5 per cent level test is:

$$\begin{array}{ccccc} A & C & B & E & D \\ 2 & 5 & 6 & 7 & 10 \\ \cline{2-4} \end{array}$$

13. TRANSFORMATIONS IN THE ANALYSIS OF VARIANCE.

The analysis of variance is based upon certain underlying assumptions--normality, homoscedasticity, additivity, and uncorrelated errors. It is assumed that each random effect(including experimental error), has a normal (Gaussian) distribution. The normal distribution is represented graphically by a certain bell-shaped curve, and is uniquely determined by two parameters, the mean and the variance (or its square root, the standard deviation). Homoscedasticity means equal variation in all subclasses. Additivity implies that the effect of two causal factors working together is the sum of their separate effects. If there is no linear relationship between two variables, they are said to be uncorrelated. If there is no relation of any kind (linear or non-linear) between two variables, they are said to be independent of each other. Normality and lack of correlation together imply independence.*

In cases where the assumptions underlying the analysis of variance are not satisfied, it may be advisable to apply a transformation to the data before performing the analysis. If the type of population from which the data have been drawn is known, there are valid theoretical grounds for choosing the transformation to be used. If the population has a Poisson distribution (as is often true in the case of integral values obtaining by counting), the square-root transformation can be shown to be appropriate. If the population has a logarithmico-normal distribution (as in the case of sample variances), the logarithmic transformation should be used. In the case of a sample correlation coefficient r, the

*See Cramér, Harald, Mathematical Methods of Statistics, p. 311.

proper transformation is $(1/2) \log [(1+r)/(1-r)]$. For a binomial population (as in the case of percentages), there are valid reasons for using the arc sine transformation, though the probit and logit transformations have been advocated for use in this case, especially in bioassay problems. In the case of ranked data, transformation to expected normal scores is recommended. All of these transformations tend to normalize the data and homogenize the variance. Transformations which do one of these things usually (but not always) do the other also, and also tend to reduce non-additivity of the effects. The remedy for correlated errors is not so much in the direction of transformations as in the direction of proper randomization procedures.

In cases where the type of population from which the data have been drawn is not known, an effort must be made to determine objectively from the data whether or not a transformation is needed and, if so, which one. Here the criterion of normality is not as useful as one might wish. Given a set of data, it is very difficult to decide what the parent population is like, though extreme departure from normality can be detected. Until fairly recently, the principal criterion for transformation of empirical data has been heterogeneity of variance (heteroscedasticity). The usual procedure has been to consider the relation between the mean and the variance of the various subclasses, and to use the transformation which is appropriate for the theoretical distribution having the same relation between mean and variance. If the variance is proportional to the mean, as for the Poisson distribution, the square-root transformation is used. If the variance is proportional to the square of the mean, as for the logarithmico-normal distribution, the logarithmic transformation is used. During the past few years, greater attention has been paid to the additivity of effects. Two theoretical advances have given impetus to this change in emphasis. Tukey proposed that a test for non-additivity be made by separating the sum of squares for discrepance into two parts, with one degree of freedom for non-additivity and the rest of the degrees of freedom for residual. If the F-ratio formed by dividing the mean square for non-additivity by the residual mean square is large, the need for a transformation to correct for non-additivity is indicated. Box showed that substantial departures from normality and homogeneity of variance have but little effect on the overall test of significance for models involving equal subclass numbers. As knowledge of these findings has spread, the tendency has been in the direction of making a transformation, where necessary, that will help to ensure additivity, and to hope that normality and homogeneity of variance will come along as by-products. If two or more transformations are satisfactory with respect to additivity, heterogeneity of variance is used as a secondary criterion for choosing the best transformation. Thus, among those transformations which reduce non-additivity to a reasonable level, one chooses that transformation which also reduces heterogeneity as much as possible.

One transformation not mentioned so far is the reciprocal transformation. This should be considered especially when the reciprocal of the original variable is just as capable of sensible interpretation as is the original variable itself. For example one may want to analyze data on frequency instead of wave length, or on conductance instead of resistance.

For a single-classification experiment, non-additivity is not defined. Hence one concerns himself with non-homogeneity of variance and non-normality. Neither of these has a serious effect on the overall F-test for equal subclass numbers. They may, however, be serious for unequal subclass numbers. They may also have a marked effect on individual comparisons, even for equal subclass numbers. In choosing a transformation for a single-classification experiment, one should seek to reduce heterogeneity, hoping that non-normality will simultaneously be reduced.

For the double-classification experiment with a and b levels of factors A and B, respectively, the sum of squares (or mean square, since it has but a single degree of freedom) for non-additivity is given by

$$\frac{ab\,\{\sum_{i}\sum_{j} y_{ij}\,(y_{i.}-y_{..})(y_{.j}-y_{..})\}^2}{(\text{SS for A})(\text{SS for B})},$$

where y_{ij} is the observation for the i^{th} level of factor A and the j^{th} level of factor B, and where the dot notation indicates an average over the missing subscript(s), viz.

$$y_{i.} = \sum_{j=1}^{b} y_{ij}/b,\ y_{.j} = \sum_{i=1}^{a} y_{ij}/a, y_{..} = \sum_{i=1}^{a}\sum_{b=1}^{j} y_{ij}/ab.$$

For the triple-classification experiment with a, b and c levels of factors A, B and C, respectively, the sum of squares (or mean square) for non-additivity is given by

$$\frac{(abc)^2 \{\sum_i \sum_j \sum_k y_{ijk}(y_{i..} - y_{...})(y_{.j.} - y_{...})(y_{..k} - y_{...})\}^2}{(\text{SS for A})(\text{SS for B})(\text{SS for C})},$$

where y_{ijk} is the observation for the i^{th} level of factor A, the j^{th} level of factor B, and the k^{th} level of factor C, and where the dot notation indicates an average over the missing subscript(s). Thus

$$y_{i..} = \sum_{j=1}^{b} \sum_{k=1}^{c} y_{ijk}/bc, \quad y_{.j.} = \sum_{i=1}^{a} \sum_{k=1}^{c} y_{ijk}/ac, \quad y_{..k} = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ijk}/ab,$$

$$y_{...} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} y_{ijk}/abc.$$

As an example of the use of transformations in the analysis of variance, consider an experiment involving the measurement of the electrical resistance of propeller blades recently conducted by the Propeller Laboratory of WADC. Eight operators individually measured the resistance of each of four propeller blades with each of two instruments, a 500-volt megger and a 1000-volt megger. The order of the 64 measurements was randomized. The data are given in Table 13.1.

Table 13.1

Resistance of Propeller Blades (megohms)

	Voltage of				Operator				
Blade	Instrument	1	2	3	4	5	6	7	8
1	500	7.50	8.00	8.10	8.00	7.50	8.00	8.00	8.00
	1000	6.00	5.00	6.10	5.75	6.00	6.00	5.80	5.00
2	500	0.30	0.30	0.31	0.31	0.30	0.31	0.30	0.30
	1000	0.30	0.30	0.30	0.30	0.40	0.35	0.35	0.30
3	500	0.31	0.31	0.30	0.40	0.35	0.33	0.32	0.31
	1000	0.30	0.45	0.30	0.30	0.35	0.35	0.35	0.30
4	500	35.00	32.00	30.00	32.00	32.00	32.00	35.00	33.00
	1000	27.00	25.00	31.00	24.00	25.00	25.00	26.00	25.00

For a three-factor experiment (without nesting), the sums of squares are most easily computed with the aid of three auxiliary two-way tables in which each entry represents, for particular levels of two-factors, the sum over all levels of the other factor. The three two-way tables constructed from the data of Table 13. 1 are Tables 13. 2, 13. 3 and 13. 4.

Table 13. 2

Voltage of	Blade				Sum
Instrument	1	2	3	4	
500	63. 10	2. 43	2. 63	261. 00	329. 16
1000	45. 65	2. 60	2. 70	208. 00	258. 95
Sum	108. 75	5. 03	5. 33	469. 00	588. 11

Table 13. 3

Blade	Operator								Sum
	1	2	3	4	5	6	7	8	
1	13. 50	13. 00	14. 20	13. 75	13. 50	14. 00	13. 80	13. 00	108. 75
2	0. 60	0. 60	0. 61	0. 61	0. 70	0. 66	0. 65	0. 60	5. 03
3	0. 61	0. 76	0. 60	0. 70	0. 70	0. 68	0. 67	0. 61	5. 33
4	62. 00	57. 00	61. 00	56. 00	57. 00	57. 00	61. 00	58. 00	469. 00
Sum	76. 71	71. 36	76. 41	71. 06	71. 90	72. 34	76. 12	72. 21	588. 11

Table 13. 4

Voltage of	Operator								Sum
Instrument	1	2	3	4	5	6	7	8	
500	43. 11	40. 61	38. 71	40. 71	40. 15	40. 64	43. 62	41. 61	329. 16
1000	33. 60	30. 75	37. 70	30. 35	31. 75	31. 70	32. 50	30. 60	258. 95
Sum	76. 71	71. 36	76. 41	71. 06	71. 90	72. 34	76. 12	72. 21	588. 11

The total sum of squares is given by

$$(7.50)^2 + \cdots + (25.00)^2 - \left(\frac{588.11^2}{64}\right) = 9336.1465$$

The sum of squares for blades is

$$\frac{(108.75)^2 + \cdots + (469.00)^2}{16} - \frac{(588.11)^2}{64} = 9085.8081$$

The sum of squares for instruments is

$$\frac{(329.16)^2+(258.95)^2}{32} - \frac{(588.11)^2}{64} = 77.0226$$

The sum of squares for operators is

$$\frac{(76.71)^2+\cdots+(72.21)^2}{8} - \frac{(588.11)^2}{64} = 5.2173$$

Remembering that each entry in Table 13.2 is the sum of eight observations, one finds the total sum of squares for this table to be

$$\frac{(63.10)^2+\cdots+(208.00)^2}{8} - \frac{(588.11)^2}{64} = 9280.4041$$

The sum of squares for the interaction between blades and instruments is then found by subtraction to be

$$9280.4041-9085.8081-77.0226 = 117.5734$$

Remembering that each entry in Table 13.3 is the sum of two observations, one finds the total sum of squares for this table to be

$$\frac{(13.50)^2+\cdots+(58.00)^2}{2} - \frac{(588.11)^2}{64} = 9105.4226$$

The sum of squares for the interaction between blades and operators is then found by subtraction to be

$$9105.4226-9085.8081-5.2173 = 14.3972$$

Remembering that each entry in Table 13.4 is the sum of four observations, one finds the total sum of squares for this table to be

$$\frac{(43.11)^2+\cdots+(30.60)^2}{4} - \frac{(588.11)^2}{64} = 91.6382$$

The sum of squares for the interaction between instruments and operators is then found by subtraction to be

$$91.6382-77.0226-5.2173 = 9.3983$$

Since the experiment is unreplicated, the three-factor interaction and the experimental error cannot be separated. Their sum will be called the three-way discrepance. The sum of squares for the three-way discrepance is found by subtraction to be

$$9336.1465-9085.8081-77.0226-5.2173-117.5734-14.3972-9.3983 = 26.7296$$

The analysis of variance and the resulting F-tests are given in Table 13.5. Since blades and instruments are assumed to be fixed and operators random, all effects not involving operators are tested by their interaction with operators, while all factors involving operators, in the absence of a separate error mean square, are tested by the three-way discrepance.

Table 13.5

Analysis of Variance for Data of Table 13.1.

Source of Variation	SS	DF	MS	F
Blades (B)	9085.8081	3	3028.6027	4417
Instruments (I)	77.0226	1	77.0226	57.37
Operators (O)	5.2173	7	0.7453	0.59
BxI	117.5734	3	39.1911	30.79
BxO	14.3972	21	0.6856	0.54
IxO	9.3983	7	1.3426	1.05
Discrepance	26.7296	21	1.2728	
Total	9336.1465	63		

The F-ratio for blades (4417) is very highly significant ($F_{.01} = 4.87$ for 3 and 21 degrees of freedom). The F-ratio for instruments (57.37) is also quite highly significant ($F_{.01} = 12.25$ for 1 and 7 degrees of freedom), as is the F-ratio for interaction between blades and instruments (30.79 as compared with $F_{.01} = 4.87$ for 3 and 21 degrees of freedom). The F-ratios for operators and for their interactions with blades and with instruments are non-significant, which indicates that differences in operators produce little if any effect, either alone or in conjunction with other factors.

The question now arises as to whether the data of Table 13.1 (or more properly the population(s) from which they come) satisfy the basic assumptions underlying the analysis of variance. Even a casual glance at the data is enough to convince an unbiased observer that the answer to this question is almost certainly in the negative. As one indication of this, consider the range of the observations for the various blades, which are given in Table 13.6, along with the range after applying three commonly used transformations.

Table 13.6

Range of Observations

Blade	Original Data	Square Roots	Logarithms	Reciprocals
1	3.10	0.61	0.21	.077
2	0.10	0.08	0.12	.833
3	0.15	0.12	0.17	1.111
4	11.00	1.01	0.16	.015

The range for the original data shows a high degree of heterogeneity. This has been reduced somewhat in the case of the square roots, but a more powerful transformation appears to be needed. The reciprocal transformation, on the other hand, is easily seen to be too powerful. The range of the logarithms shows no more heterogeneity than one might expect by chance. If one assumes that the standard deviation is proportional to the range, the Cochran test shows no significant heterogeneity. Thus the logarithmic transformation appears to be appropriate from the standpoint of homogeneity of variance.

The need for a transformation can also be demonstrated by applying Tukey's test of non-additivity. Results of this test applied to the original data and to the logarithms of the original data will be shown in Table 13.9. First, however, it is necessary to transform the data and analyze the transformed data. The common logarithms (rounded to two decimal places) of the original data are shown in Table 13.7.

Table 13.7

Logarithms of Original Data (Table 13.1)

Blade	Voltage Instrument	1	2	3	4	5	6	7	8
1	500	1.88	1.90	1.91	1.90	1.88	1.90	1.90	1.90
	1000	1.78	1.70	1.79	1.76	1.78	1.78	1.76	1.70
2	500	0.48	0.48	0.49	0.49	0.48	0.49	0.48	0.48
	1000	0.48	0.48	0.48	0.48	0.60	0.54	0.54	0.48
3	500	0.49	0.49	0.48	0.60	0.54	0.52	0.51	0.49
	1000	0.48	0.65	0.48	0.48	0.54	0.54	0.54	0.48
4	500	2.54	2.51	2.48	2.51	2.51	2.51	2.54	2.52
	1000	2.43	2.40	2.49	2.38	2.40	2.40	2.41	2.40

The transformed data in Table 13.7 are analyzed with the aid of three auxiliary two-way tables (not shown here) analogous to Tables 13.2, 13.3 and 13.4 for the original data. The mode of computing the sums of squares is the same as for the original data. The analysis of variance and the resulting F-tests are given in Table 13.8. As in Table 13.5, effects not involving operators are tested by their interaction with operators, and all effects involving operators are tested by discrepance.

Table 13.8

Analysis of Variance for Transformed Data (Table 13.7)

Source of Variation	SS	DF	MS	F
Blades (B)	46.1478	3	15.3826	13,984
Instruments (I)	0.0425	1	0.0425	30.36
Operators (O)	0.0065	7	0.0009	0.60
B x I	0.0799	3	0.0266	17.73
B x O	0.0230	21	0.0011	0.73
I x O	0.0101	7	0.0014	0.93
Discrepance	0.0310	21	0.0015	
Total	46.3408	63		

The same effects are found to be significant as in the analysis of the original data, but it should not be concluded from this fact that no transformation is necessary. The F-ratio for blades is more than three times as large as for the original data, and the F-ratios for instruments and for interaction between blades and instruments are little more than half as large. If they had been near the critical values, differences in the conclusions might well have resulted. Furthermore, if individual comparisons are to be made following the analysis of variance, a pooled estimate s^2 on a scale where the dispersions of the groups vary as much as in the original data (see Table 13. 6) is not appropriate. The pooled $s^2 = 0.6856$ used for testing blade differences in Table 13. 5 is much too large for comparing blades 2 and 3, which have small dispersions, and much too small for comparing blades 1 and 4, which have large dispersions.

Now consider the question of additivity, as measured by Tukey's test. The sums of squares for discrepance in Tables 13. 5 and 13. 8 can each be broken down or decomposed into a sum of squares (or mean square, since it has but a single degree of freedom) for non-additivity and a residual sum of squares. The sums of squares for non-additivity in the three-way tables (Table 13. 1 and 13. 7) are computed by use of the formula for a three-factor experiment given on page 47. Similarly, the sums of squares for BxI, BxO, and IxO can be decomposed. The sums of squares for non-additivity in the auxiliary two-way tables (Tables 13. 2-13. 4 and their analogues formed from Table 13. 7) are computed by use of the formula for a two-factor experiment given on page 46, with the proper adjustment for the number of observations upon which each item in the table is based. In each case the test for non-additivity is made by dividing the mean square for non-additivity by the residual mean square and comparing the result with the 5 per cent and 1 per cent points of the F-distribution with 1 degree of freedom for the numerator and the proper number for the denominator. The results of applying these tests to the original data and to the transformed data are given in Table 13. 9. Two of the F-ratios for non-additivity are significant (one at the 1 per cent level and the other at the 5 per cent level) for the original data. None of the F-ratios for the transformed data is significant. Hence, from the point of view of both homogeneity of variance and additivity of effects, the logarithmic transformation is appropriate for these data.

Now consider the application of the Newman-Keuls procedure to make individual comparisons of the blade means and the interactions between blades and instruments. Since there are only two instruments, the F-test tells the whole story about them, and the main effect of operators and interactions involving operators are not significant, so no individual comparisons are made for these effects. The proper scale for making the individual comparisons is the logarithmic scale, but they will be made also on the original scale in order to demonstrate the differences in the results.

Table 13. 9

Results of Tukey's Test on Non-additivity

Interaction	F for non-additivity		Critical values	
Decomposed	Original data	Logarithms	$F_{.05}$	$F_{.01}$
Discrepance	2. 28	0. 80	4. 35	8. 10
BxI	259.	8. 44	18. 51	98. 49
BxO	4. 66	1. 08	4. 35	8. 10
IxO	1. 36	4. 10	5. 59	12. 25

Since the interaction between blades and operators (BxO) is used in testing the effect of blades, the mean square for BxO will be taken as the estimate s^2 of the variance within blades, which is used in computing the standard error of the mean for blades. Thus the standard error of this mean, $s_{\bar{x}}$, is given by

$$s_{\bar{x}} = \sqrt{0.6856/16} = \sqrt{.04285} = 0.207 \quad \text{(Original data)}$$

$$s_{\bar{x}} = \sqrt{0.0011/16} = \sqrt{.0000688} = 0.00829 \text{ (Logarithms)}$$

Multiplying these values by the 5 per cent and 1 per cent critical values for the studentized range of p observations (p = 2, 3, 4) with 21 degrees of freedom for s^2, one finds the critical differences between blade means required for significance by the Newman-Keuls procedure. These critical differences and the resulting tests are shown in Tables 13. 10 (original data) and Table 13. 11 (logarithms). Both analyses show that all pairs of blade means differ significantly except the means for blades 2 and 3. Actually, the difference between blades 2 and 3 is just short of significance at the 5 per cent level (0. 023 as compared with $ISD_{.05}$ = 0. 024 on the logarithmic scale). The principal difficulty with the analysis on the original scale is that the pooled estimate of variances which are heterogeneous is much too large to use in testing the significance of the difference between blades 2 and 3, both of which have small variances, and hence greatly underestimates the significance of this difference.

Table 13.10

Individual Comparisons of Blade Means (Original Data)

Significance Level (α)	Critical Difference ISD_α			Mean for Blade r			
	p = 2	p = 3	p = 4	r = 2	r = 3	r = 1	r = 4
5%	0.61	0.74	0.82	0.31	0.33	6.80	29.31
1%	0.83	0.95	1.03	0.31	0.33	6.80	29.31

Table 13.11

Individual Comparisons of Blade Means (Logarithms)

Significance Level (α)	Critical Difference ISD_α			Mean for Blade r			
	p = 2	p = 3	p = 4	r = 2	r = 3	r = 1	r = 4
5%	0.024	0.030	0.033	0.497	0.520	1.826	2.465
1%	0.033	0.038	0.041	0.497	0.520	1.826	2.465

The interaction of blades r and s (r, s = 1, 2, 3, 4; r≠ s) with instruments 1 and 2 will be designated by

$$(I_1 - I_2)(B_r - B_s) = (I_1 - I_2)B_r - (I_1 - I_2)B_s \quad .$$

The interaction element $(I_1 - I_2)\, B_r$ is defined as the mean of the readings on blade r using instrument 1 minus the mean of the readings on blade r using instrument 2. For example, on the original scale,

$$(I_1 - I_2)B_4 = (261.00 - 208.00)/8 = 6.625.$$

Since discrepance is used in testing the interaction between blades and instruments, the mean square for discrepance will be taken as the estimate s^2 of the variance within blade-instrument combinations, which is used in computing the standard error of the interaction element. This standard error, $s_{\bar{d}}$, is the standard error of the difference between the means of two samples each of size 8. Hence the standard error is given by

$$s_{\bar{d}} = \sqrt{2(1.2728)/8} = \sqrt{0.3182} = 0.564 \quad \text{(Original data)}$$

$$s_{\bar{d}} = \sqrt{2(0.0015)/8} = \sqrt{0.000375} = 0.0194 \quad \text{(Logarithms)}$$

Multiplying these values by the 5 per cent and 1 per cent critical values for the studentized range of p observations (p = 2, 3, 4) with 21 degrees of freedom for s^2, one finds the critical interactions of blades with instruments required for significance by the Newman-Keuls procedure. These critical interactions and the resulting tests are shown in Table 13. 12 (original data) and Table 13. 13 (logarithms).

Table 13. 12
Individual Comparisons of Blade-Instrument Interaction Elements
(Original Data)

Significance Level (α)	Critical Difference ISD_α			Interaction Element $(I_1 - I_2)B_r$			
	p = 2	p = 3	p = 4	r = 2	r = 3	r = 1	r = 4
5%	1.658	2.008	2.222	-0.021	-0.009	2.181	6.625
1%	2.256	2.600	2.814	-0.021	-0.009	2.181	3.312

Table 13. 13

Individual Comparisons of Blade-Instrument Interaction Elements
(Logarithms)

Significance Level (α),	Critical Difference ISD_α			Interaction Element $(I_1 - I_2)B_r$			
	p = 2	p = 3	p = 4	r = 2	r = 3	r = 4	r = 1
5%	.0568	.0688	.0762	-.0262	-.0088	.1012	.1400
1%	.0774	.0892	.0966	-.0262	-.0088	.1012	.1400

The analysis of the original data shows no significant difference at the 1 per cent level between the interaction element of instruments with blade 1 and the interaction elements of instruments with blades 2 and 3; that is, at the 1 per cent level, it shows no significant interaction of instruments with blades 1 and 2 or with blades 1 and 3. The true perspective on this situation is given by the analysis of the transformed data, where the resistance of blades 1 and 4 is found to be much higher when measured by the 1000-volt megger than when measured by the 500-volt megger, while the resistance of blades 2 and 3 is slightly lower when measured by the 1000-volt megger than when measured by the 500-volt megger. The difference between the resistances found by the two instruments is greater for blade 1 than for blade 4 when measured on the logarithmic scale, corresponding to the fact that the ratio of the resistances is greater for blade 1 than for blade 4 on the original scale. (The difference of the logarithms of two numbers is the logarithm of their ratio.)

REFERENCES

Anderson, R. L.; and Bancroft, T. A., Statistical Theory in Research, 1952. New York, McGraw-Hill Book Company.

Anscombe, F. L., The validity of comparative experiments, Journal of the Royal Statistical Society, Series A, 111(1948), pp. 181-200.

Bancroft, T. A., On biases in estimation due to the use of preliminary tests of significance. Annals of Mathematical Statistics, 15(1944).

Barnes, Benjamin A., The analysis of variance: a graphical representation of a statistical concept. Journal of the American Statistical Association, 50(1955), pp. 1064-1121.

Bartlett, M. S., Square-root transformation in analysis of variance. Journal of the Royal Statistical Society, Supplement, 3(1936), pp. 68ff.

Bartlett, M. S., The use of transformations, Biometrics, 3(1947), pp. 39-52.

Bartlett, M. S.; and Kendall, D. G., The statistical analysis of variance-heterogeneity and the logarithmic transformation, Journal of the Royal Statistical Society, Supplement, 7(1946), pp. 128ff.

Bennett, Carl A., Applications of tests for randomness. Industrial and Engineering Chemistry, 43(1951), p. 2063.

Bennett, Carl A.; and Franklin, Norman L., Statistical Analysis in Chemistry and the Chemical Industry, 1954. New York, John Wiley and Sons, Inc.

Berkson, J., Tests of significance considered as evidence. Journal of the American Statistical Association, 37(1942) pp. 324-335.

Berkson, Joseph, Application of the logistic function to bio-assay. Journal of the American Statistical Association, 39(1944), pp. 357ff.

Bliss, C.I.; and Calhoun, D.W., An Outline of Biometry, 1954. New Haven, Yale Co-Operative Corporation.

Box, G.E.P., Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, Annals of Mathematical Statistics, 25(1954), pp. 290-302.

Box, G.E.P., Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 25(1954), pp. 484-498.

Bozivich, Helen; Bancroft, T.A.; Hartley, H.O.; and Huntsberger, David, V., Analysis of Variance: Preliminary Tests, Pooling, and Linear Models, WADC Technical Report 55-244, Volume I, Preliminary Tests of Significance and Pooling Procedures for Certain Incompletely Specified Models.

Brownlee, K.A., Industrial Experimentation, Third American Edition, 1949. Brooklyn, Chemical Publishing Company.

Brownlee, K.A., Experiments with many factors. Chemical Engineering Progress, 49(1953), pp. 617-621.

Claringbold, P.J.; Biggers, J.D.; and Emmens, C.W., The angular transformation in quantal analysis. Biometrics, 9(1953), pp. 475ff.

Cochran, W.G., Some difficulties in the statistical analysis of replicated experiments. Empire Journal of Experimental Agriculture, 6(1938), pp. 157-175.

Cochran, W.G., The analysis of variance when experimental errors follow the Poisson or binomial laws. Annals of Mathematical Statistics, 11(1940), pp. 335ff.

Cochran, W.G., Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 3(1947), pp. 22-38.

Cochran, W.G., Testing a linear relation among variances. Biometrics, 7(1951), pp. 17-32.

Cochran, William G.; and Cox, Gertrude M., *Experimental Designs*, 1950. New York, John Wiley and Sons, Inc.

Cornfield, Jerome, On samples from finite populations. *Journal of the American Statistical Association*, 39(1944), pp. 236-239.

Cornfield, Jerome, The two-way classification in the analysis of variance from the point of view of urn sampling. Mimeographed notes, 1953.

Cramér, H., *Mathematical Methods of Statistics*, 1946. Princeton, Princeton University Press.

Crump, S. L., The present status of variance component analysis. *Biometrics*, 7(1951), pp. 1-16.

Curtiss, J. H., On transformations used in analysis of variance. *Annals of Mathematical Statistics*, 14(1943), pp. 10ff.

Davies, Owen L. (editor); et al, *Design and Analysis of Industrial Experiments*, 1954. New York, Hafner Publishing Company.

Deming, W. E., *Some Theory of Sampling*, 1950. New York, John Wiley and Sons, Inc.

Dixon, Wilfred J.; and Massey, Frank J., Jr., *Introduction to Statistical Analysis*, 1951. New York, McGraw-Hill Book Company, Inc.

Duncan, D. B., A significance test for differences between ranked treatments in an analysis of variance, *Virginia Journal of Science*, 2(1951), pp. 171-189.

Duncan, D. B., On the properties of the multiple comparisons test, *Virginia Journal of Science*, 3(1952) pp. 49-67.

Duncan, David B., Multiple range and multiple F tests. *Biometrics*, 11(1955), pp. 1-42.

Edwards, Allen L., *Experimental Design in Psychological Research*, 1950. New York, Rinehart & Company, Inc.

Eisenhart, Churchill, The assumptions underlying the analysis of variance. Biometrics, 3(1947), pp. 1-21.

Finney, D. J., On the distribution of a variate whose logarithm is normally distributed. Journal of the Royal Statistical Society, Supplement, 7(1949), pp. 155-161.

Finney, D. J., Probit Analysis, 1952. Cambridge, University Press.

Fisher, R. A., Statistical Methods for Research Workers. Edinburgh and London, Oliver and Boyd.

Fisher, R. A., The Design of Experiments. Edinburgh and London, Oliver and Boyd.

Fisher, R. A., The arrangement of field experiments. Journal of Mining, Agriculture, and Engineering, 33(1926), pp. 503-513.

Fisher, R. A., The analysis of variance with various binomial transformations. Biometrics, 10(1954), pp. 130-139.

Freeman, H. A., Industrial Statistics, 1942. New York, John Wiley & Sons, Inc.

Greenberg, B. G., Why randomize? Biometrics 7(1951), pp. 309-322.

Grundy, P. M.; and Healy, M. J. R., Restricted randomization and quasi-latin squares, Journal of the Royal Statistical Society, 12(1950), pp. 286-291.

Hald, A., Statistical Theory with Engineering Applications, 1952. New York, John Wiley & Sons, Inc.

Harter, H. Leon, Error rates and sample sizes in multiple comparisons. Unpublished paper presented at Montreal meeting of American Statistical Association, 1954.

Harter, H. Leon, Error rates and sample sizes for multiple range tests. Unpublished paper presented at Ann Arbor meeting of Institute of Mathematical Statistics, 1955.

Harter, H. Leon; and Lum, Mary D., Partially Hierarchal Models in the Analysis of Variance, WADC Technical Report 55-23. Wright-Patterson Air Force Base.

Hartley, H. O., The use of range in analysis of variance. *Biometrika,* 37(1950), pp. 271-280.

Hartley, H. O., Some recent developments in analysis of variance. *Communications on Pure and Applied Mathematics*, 8(1955), pp. 47-72.

Hoel, Paul G., *Introduction to Mathematical Statistics,* 2nd edition, 1954. New York, John Wiley and Sons, Inc.

Horton, H. Burke, *Table of 105,000 Random Decimal Digits*, 1949, Washington, Interstate Commerce Commission, Bureau of Transport Economics.

Huntsberger, D. V., A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics*, 26(1955), pp. 734-743.

Irwin, J. O., Mathematical theorems involved in the analysis of variance. *Journal of the Royal Statistical Society*, 94(1931), pp. 284-.

Kempthorne, Oscar, *The Design and Analysis of Experiments*, 1952. New York, John Wiley & Sons, Inc.

Kempthorne, Oscar, The randomization theory of experimental inference. *Journal of the American Statistical Association,* 50(1955), pp. 946-967.

Kendall, M. G., *The Advanced Theory of Statistics,* Volumes I and II, 1948. London, Charles Griffin & Company, Limited.

Kendall, M. G.; and Smith, B. Babington, *Tables of Random Sampling Numbers,* Tracts for Computers, No. 24 (1939), Cambridge, University Press.

Kenney, J. F.; and Keeping, E. S., *Mathematics of Statistics,* Part Two, 2nd edition, 1951. New York, D. Van Nostrand Company, Inc.

Keuls, M., The use of studentized range in connection with an analysis of variance. *Euphytica*, 1(1952), pp. 112-122.

Lacey, Oliver L., *Statistical Methods in Experimentation: An Introduction.* 1953. New York, The Macmillan Company.

Lindquist, E. F., *Statistical Analysis in Educational Research,* 1940. Boston, Houghton Mifflin Company.

Lum, Mary D., Rules for determining error terms in hierarchal and partially hierarchal models. Unpublished paper presented at Iowa City meeting of Institute of Mathematical Statistics, 1954.

Mann, H. B., *Analysis and Design of Experiments,* 1949. New York, Dover Publications, Inc.

Mentzer, Eldo G., Tests by the Analysis of Variance, WADC Technical Report 53-23. Wright-Patterson Air Force Base.

Merrington, M.; and Thompson, C. M., Tables of percentage points of the inverted beta (F) distribution. *Biometrika,* 33(1943), pp. 73-88.

Mood, Alexander McFarlane, *Introduction to the Theory of Statistics,* 1950. New York, McGraw-Hill Book Company, Inc.

Mosteller, Frederick, On pooling data. *Journal of the American Statistical Association,* 43(1948), pp. 231-242.

Newman, D., The distribution of range in samples from a normal population expressed in terms of an independent estimate of standard deviation. *Biometrika,* 31(1939), pp. 20ff.

Neyman, J.; and Pearson, E. S., On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions,* Series A, 231(1933), pp. 289-337.

Olds, Edwin G.; and Severo, Norman C., A Comparison of Tests on the Mean of a Logarithmico-Normal Distribution with Known Variance, WADC Technical Note 55-249.

Paull, A. E., On a preliminary test for pooling mean squares in the analysis of variance. *Annals of Mathematical Statistics,* 21(1950), pp 539-556.

Pearson, E. S.; and Hartley, H. O., *Biometrika Tables for Statisticians,* Volume 1, 1954. Cambridge, University Press.

Pitman, E. J. G., Significance tests which may be applied to samples from any population. III. The analysis of variance test. *Biometrika*, 29(1937), pp. 322-335.

Quenouille, M. H., *The Design and Analysis of Experiment*, 1953. New York, Hafner Publishing Company.

Rand Corporation, *A Million Random Digits*, Glencoe, Illinois, The Free Press [Some of these digits are published in the *Journal of the American Statistical Association*, Volumes 47-49 (1952-54)].

Rao, C.R., *Advanced Statistical Methods in Biometric Research*, 1954. New York, John Wiley & Sons, Inc.

Satterthwaite, F.E., Synthesis of variance. *Psychometrika*. 6(1941), pp. 309-316.

Satterthwaite, F. E., An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(1946), pp. 110-114.

Scheffe', Henry, A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1953), pp. 87-104.

Scheffe', H., Statistical methods for evaluation of several sets of constants and several sources of variability. *Chemical Engineering Progress*, 50(1954), pp. 200-205.

Scheffe', Henry, Alternate models in the analysis of variance. Unpublished paper presented at Ann Arbor meeting of Institute of Mathematical Statistics, 1955.

Schultz, E. F., Rules of thumb for determining expectations of mean squares in analysis of variance. *Biometrics*, 11(1955), pp. 123-135.

Smith, H. F., Variance components, finite populations, and experimental inference. Institute of Statistics Mimeo Series No. 135, North Carolina State College, 1955.

Snedecor, George W., *Statistical Methods*, 4th edition, 1946. Ames, Iowa. The Iowa State College Press.

Tippett, L. H. C., Random Sampling Numbers, Tracts for Computers, No. 15(1927), Cambridge, University Press.

Tippett, L. H. C., Technological Applications of Statistics, 1950. New York, John Wiley & Sons, Inc.

Tukey, J. W., One degree of freedom for non-additivity. Biometrics, 5.(1949), pp. 232-242.

Tukey, J. W., Interaction in a row by column design. Memorandum Report 18, Princeton University, 1949.

Tukey, J. W., The problem of multiple comparisons. Unpublished memorandum, 1953.

Vaurio, V. W.; and Daniel, Cuthbert, Evaluation of several sets of constants and several sources of variability. Chemical Engineering Progress, 50(1954), pp. 81-86.

Villars, Donald Statler, Statistical Design and Analysis of Experiments for Development Research, 1951. Dubuque, Iowa, Brown.

Wilk, M. B., The randomization analysis of a generalized randomized block design. Biometrika, 42(1955), pp. 70-79.

Wilk, M. B.; and Kempthorne, O., Fixed, mixed, and random models. Journal of the American Statistical Association, 50(1955), pp. 1144-1167.

Wilk, M. B.; and Kempthorne, O., Analysis of Variance: Preliminary Tests, Pooling, and Linear Models, WADC Technical Report 55-244, Volume II, Derived Linear Models and Their Use in the Analysis of Randomized Experiments.

Yates, F. Incomplete Latin squares. Journal of Agricultural Science, 26 (1936), pp. 301-315.

Youden, W. J., Statistical Methods for Chemists, 1951. New York, John Wiley & Sons, Inc.

ADDENDA

Anscombe, F.L., and Tukey, J.W., The Criticism of Transformations. Unpublished paper presented at meeting of American Statistical Association and Biometrics Society, Montreal, 1954.

Harter, H. Leon, Fractionally Replicated Designs for Testing Titanium Alloys, WADC Technical Note 55-14.

McNemar, Quinn, Psychological Statistics, 2nd edition, 1955. New York, John Wiley and Sons, Inc.